UNIVERSITY OF CALIFORNIA,
IRVINE

Frameworks for Personalized Fairness and Privacy-aware Human-in-the-Loop Systems

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Engineering

by

Tianyu Zhao

Dissertation Committee:
Professor Salma Elmalaki, Chair
Professor Athina Markopoulou
Professor Yasser Shoukry

2026

DEDICATION

Dedicated to my parents and my partner, whose love, patience, and encouragement made this work possible.

Dedicated to my younger self, who decided to pursue this degree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I am grateful to my advisor, Professor Salma Elmalaki, who directed and supervised the research that forms the basis of this dissertation. Her work on human-centered, fairness- and privacy-aware computing shaped every chapter, as did her mentorship in the Pervasive Autonomy Lab at the University of California, Irvine. I also thank my committee members, Professor Athina Markopoulou and Professor Yasser Shoukry, for their time, feedback, and guidance.

I thank my collaborators, whose contributions appear throughout these chapters: Mojtaba Taherisadr (FAIRO), Mahmoud Srewa (FinP, PluralLLM, and the preference-aggregation study), and Siqi Li (PACIFIC). I also thank the rest of the Pervasive Autonomy Lab for years of discussion and critique.

This research was supported by the U.S. National Science Foundation (NSF) under awards 2105084 (CRII: CPS: Society-in-the-Loop Personalized Computing) and 2339266 (CAREER: Psychology-aware Human-in-the-Loop Cyber-Physical-System (HCPS): Methodologies, Algorithms, and Deployment). The PACIFIC work was additionally partially supported by the UCI ProperAI Institute, an Engineering+Society Institute funded as part of a generous gift from Susan and Henry Samueli.

Portions of this dissertation reprint previously published material. Chapter 2 is a reprint of material previously published by IEEE. The IEEE does not require a formal reuse license for use in a thesis. © 2024 IEEE. Reprinted, with permission, from T. Zhao and S. Elmalaki, "F*in*A: Fairness of Adverse Effects in Decision-Making of Human-Cyber-Physical-System," 2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS), May 2024. © 2024 IEEE. Reprinted, with permission, from T. Zhao, M. Taherisadr, and S. Elmalaki, "FAIRO: Fairness-aware Sequential Decision Making for Human-in-the-Loop CPS," 2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS), May 2024. The coauthor Professor Salma Elmalaki directed and supervised the research that forms the basis for this chapter. IEEE has stated that University Microfilms and/or ProQuest may supply single copies of the dissertation. Chapter 3 reprints material that appeared in the Proceedings on Privacy Enhancing Technologies (PoPETs) under a Creative Commons Attribution 4.0 license. Chapter 4 reprints material that appeared in the 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving in COLM 2026. Chapter 5 reprints material that appeared in the Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems (HumanSys), published by ACM, and in the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle. The coauthor Professor Salma Elmalaki listed in these publications directed and supervised the research that forms the basis for this dissertation.

VITA

Tianyu Zhao

EDUCATION

Doctor of Philosophy in Computer Engineering **2026**
University of California, Irvine *Irvine, California*
Dissertation: Frameworks for Personalized Fairness and Privacy-aware Human-in-the-Loop Systems
Advisor: Professor Salma Elmalaki

Master of Science in Computer Engineering **2021**
University of California, Irvine *Irvine, California*

Bachelor of Engineering in Information Engineering **2019**
South China University of Technology *Guangzhou, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2021–2026**
Pervasive Autonomy Lab, University of California, Irvine *Irvine, California*
Research on fairness, privacy, and human alignment in AI-driven human-cyber-physical systems, including sequential decision-making, federated learning, and large language models.

TEACHING EXPERIENCE

Teaching Assistant **2022–2024**
EECS 112L Digital Computer Organization Lab *Irvine, California*
University of California, Irvine. Fall 2022, Winter 2023, Fall 2023, and Winter 2024. Helped organize the first FPGA laboratory offering in the EECS department.

Undergraduate Student Mentor **2023**
Pervasive Autonomy Lab, University of California, Irvine *Irvine, California*
Mentored undergraduate researcher Yixin Zhang on a project that led to a first-author workshop paper at DESTION 2024.

HONORS AND AWARDS

First Place, Software & Algorithms Track, GEECS Annual Technology Showcase (GEECSANTS) **2025**
UC Irvine Graduate Division Completion Fellowship **2025**
UC Irvine EECS Department Graduate Fellowship **2021**

REFEREED CONFERENCE PUBLICATIONS

FinP: Fairness-in-Privacy in Federated Learning by Addressing Disparities in Privacy Risk **2026**
Tianyu Zhao, Mahmoud Srewa, and Salma Elmalaki. 26th Privacy Enhancing Technologies Symposium (PETS), Calgary, Canada.

FAIRO: Fairness-aware Adaptation in Sequential-Decision Making for Human-in-the-Loop Systems **2024**
Tianyu Zhao, Mojtaba Taherisadr, and Salma Elmalaki. 15th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), Hong Kong, China.

FinA: Fairness of Adverse Effects in Decision-Making of Human-Cyber-Physical-System **2024**
Tianyu Zhao and Salma Elmalaki. 15th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), Hong Kong, China.

REFEREED WORKSHOP PUBLICATIONS

PACIFIC: Can LLMs Discern the Psychometric Traits Influencing Your Preferences? Evaluating Personality-Driven Preference Alignment in LLMs **2026**
Tianyu Zhao*, Siqi Li*, Yasser Shoukry, and Salma Elmalaki. 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving, Conference on Language Modeling (COLM). Equal contribution with Siqi Li. Extension under review.

A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs **2025**
Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. NeurIPS Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling, San Diego, California. Extension under review.

PluralLLM: Pluralistic Alignment in LLMs via Federated Learning **2025**
Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems (HumanSys), Irvine, California.

Towards Behavioral-aware Crowd Management System **2024**
Yixin Zhang, Tianyu Zhao, and Salma Elmalaki. 6th Workshop on Design Automation for CPS and IoT (DESTION), Hong Kong, China.

ABSTRACT OF THE DISSERTATION

Frameworks for Personalized Fairness and Privacy-aware Human-in-the-Loop Systems

By

Tianyu Zhao

Doctor of Philosophy in Computer Engineering

University of California, Irvine, 2026

Professor Salma Elmalaki, Chair

Human-in-the-loop (HITL) systems, from smart environments to conversational AI assistants, increasingly couple human preferences, behavior, and feedback into the computational loop. As these systems mediate decisions that affect people, three human-centered properties cannot be separated from performance: whether the system treats people *fairly*, whether it protects their *privacy* equitably, and whether it *aligns* with what individuals and groups actually want. This dissertation develops computational frameworks that build fairness, privacy, and personalization directly into HITL systems, covering cyber-physical systems and large language models (LLMs).

The dissertation advances four contributions. First, it formalizes fairness in sequential, human-in-the-loop decision-making. *FinA* grounds fairness in the psychology of loss aversion, defining and minimizing an "adverse effect" through optimization, while *FAIRO* recasts the same goal as temporally abstracted sub-goals solved with the Options reinforcement-learning framework; together they improve perceived fairness by up to 67% over the state of the art across smart-home, water-allocation, and personalized-learning applications. Second, it treats privacy as an equity problem. *FinP* shows that privacy risk in federated learning is inequitably distributed, falling hardest on outlier clients, and enforces fairness-in-privacy through Hessian-ranked, Lipschitz-based client regularization and risk-aware server

aggregation, reducing vulnerability disparities by up to 57% with negligible utility cost. Third, it studies personalization at the level of the individual. *PACIFIC* demonstrates that stable Big-Five personality traits are a more reliable latent signal for personalization than raw conversation history, and introduces a persona-aware retriever that raises label-free preference-prediction accuracy without trait annotations. Fourth, it studies pluralistic *AI alignment* across groups: *PluralLLM* learns group preferences via federated learning as a privacy-preserving reward model, and a systematic evaluation of federated preference aggregation shows that adaptive, history-aware weighting improves fairness across groups while preserving alignment quality.

The four topics proceed in this order: fair decision making across a group over time; equitable distribution of information-leakage risk; personalization to an individual through stable personality traits; and alignment of generative models with the values of many groups. Being fair to people as a set is a different problem from being aligned to any one of them, which is why the work moves from group equity to individual personalization and then to group alignment. In all four, human experience, risk, and preference are treated as quantities to be measured, distributed, and reasoned over. The results show that fairness, privacy, and alignment can be achieved with minimal or no utility cost.

Chapter 1

Introduction

1.1 Motivation

Smart homes, health monitors, advanced driver-assistance systems, and conversational large language models (LLMs) continuously sense human state, infer human preference, and adapt their behavior in response. These *human-in-the-loop* (HITL) systems close a cognition–preference–perception cycle: the system builds a cognition of its environment, humans express preferences for how it should act, the system takes an adaptation action, and it then estimates each person's perception of that action, feeding the result back into the loop. That coupling is useful, and it also raises the stakes. A single agent's decisions may be experienced by many people at once, and a model's outputs may be tailored to one; in both cases the system allocates outcomes, risks, and representation among human beings, whether or not it was designed to.

This dissertation asks how to build computational frameworks that are fair and privacy-aware while faithfully incorporating human preference, intention, and feedback into the control or recommendation loop. The work is organized around three human-centered properties that

cannot be separated from performance. **Fairness**: when multiple people with different preferences and perceptions are affected by the same adaptation decisions, they must be treated equitably across time. **Privacy**: in multi-human and federated settings, the burden of privacy risk should be distributed fairly, so that it does not fall hardest on those already most vulnerable. **Personalization and alignment**: a system must adapt to different human preferences and personas at the individual level, and generative models must be aligned with the plural values of many groups at once. These three properties define the four topics of this work.

1.2 Background: Human-in-the-Loop Systems

Classical algorithmic fairness and privacy have most often been studied in static, one-shot settings. In applications such as pretrial risk assessment, lending, or hiring, a decision is made once from data, and fairness is a property of that single prediction. A HITL cyber-physical system (CPS) does not work that way. It is *dynamic and sequential*: an action taken now shapes future human states and future actions, human preferences vary both between people and within the same person over time, and a policy that appears fair at one moment can become discriminatory later. The same sequential character applies to privacy. In federated and multi-human environments, what matters is the *distribution* of leakage risk across participants, not a single average or worst-case figure. Personalization is sequential as well: the signal to be modeled is an evolving, noisy, and often unstated account of who a person is, not a fixed label. Every chapter that follows starts from these three characteristics: the systems are dynamic, they involve multiple humans, and they are driven by preference.

1.3 Historical Development of This Work

The research developed in the order the chapters follow. The earliest work, completed before the qualifying examination, addressed *fairness in decision-making* for multi-human CPS from

two directions: an optimization-based formulation grounded in loss aversion (FinA) and a learning-based formulation grounded in temporal abstraction (FAIRO). Both defined fairness as the minimization of disparity in human experience over time, and that definition carried into everything after it.

A system's decisions are not the only thing that affects the people in the loop; what it *leaks* matters too. That led to *privacy*, and to treating privacy risk in federated learning as an equity problem (FinP), where the question is how much information is exposed and *to whom* the exposure accrues. The work then moved from the group to the individual, asking in *personalization* whether stable psychological traits are a more reliable basis for adapting to a person than their raw interaction history (PACIFIC). Fair treatment of people as a set still leaves open whether a model is aligned to any one of them. The last topic is therefore *pluralistic AI alignment* of LLMs through federated preference learning (PluralLLM), together with a systematic evaluation of how group preferences should be aggregated.

1.4 Methods Overview

The four topics draw on a shared set of methods, adapted to each setting. Fairness is formalized through equity-based objectives: in FinA, an "adverse effect" derived from prospect theory and loss aversion, minimized by optimization at each step; in FAIRO, the fairness goal is decomposed into sub-goals over time using the Options reinforcement-learning framework, with a policy over options that repeatedly targets whoever is falling behind. Privacy is analyzed through the geometry of learning. Loss-surface curvature via the Hessian ranks client overfitting, and Lipschitz constraints on local loss curb memorization; both are closed in a loop with risk-aware server aggregation. Personalization and alignment are studied with generative models: Big-Five (OCEAN) personality traits as a compact latent abstraction of preference (PACIFIC), a persona-aware contrastive retriever that favors psychometric over semantic similarity, and federated learning with reinforcement learning from human feed-

Figure 1.1: The four research topics of this dissertation on human-in-the-loop systems (fairness, privacy, personalization, and AI alignment) and the algorithms and frameworks contributed under each.

back (RLHF) to learn and aggregate group preferences without centralizing sensitive data (PluralLLM and the preference-aggregation study). Evaluation in every chapter reports both the human-centered metric of interest (disparity of experience, disparity of privacy risk, preference-prediction accuracy, group alignment and fairness) and the utility trade-off, testing whether the human-centered gains come at the expense of task performance.

1.5 Organization of the Dissertation

This is a paper-based dissertation. Chapter 1 is the general introduction, Chapters 2–5 reproduce previously published or submitted work in full, and Chapter 6 concludes. Each of the four middle chapters follows the same pattern: an Overview section that states the problem and summarizes the approach, then the reproduced paper or papers, then a Discussion section. The chapters are ordered by topic, not by date of publication.

Figure 1.1 shows the four topics and the algorithms and frameworks developed under each. The figure also lists a line of work on fair crowd management, presented at DESTION 2024, which is part of the same research program but is not reproduced as a chapter here.

- **Chapter 2 (Fairness)** presents FinA [Zhao and Elmalaki, 2024] (Section 2.2) and FAIRO [Zhao et al., 2024] (Section 2.3), two complementary frameworks, one optimization-based and one learning-based, for fair sequential decision-making in multi-human CPS.

- **Chapter 3 (Privacy)** presents FinP [Zhao et al., 2026b] (Section 3.2), which treats privacy in federated learning as the fair distribution of privacy risk and enforces it through client-side regularization and server-side risk-aware aggregation.

- **Chapter 4 (Personalization)** presents PACIFIC [Zhao et al., 2026a] (Section 4.2), which uses Big-Five personality traits as a principled latent signal for personalizing LLM responses, together with the persona-aware retriever PiRAG.

- **Chapter 5 (AI Alignment)** presents PluralLLM [Srewa et al., 2025a] (Section 5.2), a federated approach to pluralistic alignment, and a systematic evaluation of preference-aggregation strategies for federated RLHF [Srewa et al., 2025b] (Section 5.3).

- **Chapter 6 (Conclusion)** summarizes the argument and outlines future directions.

1.6 Summary of Contributions and Results

The dissertation makes the following contributions:

- **A formal, equity-based account of fairness in sequential human-in-the-loop decisions.** FinA introduces five formulations of adverse effect (instantaneous, accumulated long-term, and perception-based) and improves perceived fairness by an average of roughly 67% over the prior state of the art; FAIRO recasts the same goal via the Options framework and improves fairness by about 35% on average across smart-home, water-resource, and personalized-learning applications, without sacrificing thermal comfort or other utility.

- **Privacy as equity.** FinP formalizes and enforces fairness-in-privacy in federated learning; it reduces privacy-risk disparity across clients by up to 57% while keeping global utility within roughly 2% of standard baselines, and it shows that enforcing privacy equity can additionally lower overall attack success, an effect uniform differential privacy does not achieve.

- **Personality traits as a compact signal for personalization.** PACIFIC shows that with clean, trait-aligned context LLMs reach near-ceiling preference-prediction accuracy (up to 99%), that the bottleneck is retrieval, not the model's reasoning, and that a persona-aware contrastive retriever raises label-free accuracy substantially over standard semantic retrieval, while surfacing systematic biases against "low"-trait personas.

- **Practical, privacy-preserving pluralistic alignment.** PluralLLM learns group preferences via federated learning as a lightweight reward model, achieving faster convergence and comparable group fairness to centralized training; the accompanying evaluation of federated preference aggregation shows that an adaptive, history-aware weighting scheme consistently improves fairness across groups while maintaining competitive alignment.

Chapter 2

Fairness in Human-in-the-Loop Decision-Making

2.1 Overview

This chapter concerns fairness in the multi-human human-in-the-loop (HITL) cyber-physical system (CPS). Several people share a physical space (rooms in a home, a shared water resource, a personalized-learning environment), and each brings a distinct, and often changing, set of preferences and perceptions. A single adaptation agent serves all of them at once, so its decisions distribute experience across people; they cannot all be satisfied at once. A thermostat policy that keeps one occupant comfortable may leave another cold. A resource-allocation policy that satisfies today's demand may systematically disadvantage the same participant tomorrow. This chapter addresses how that experience is distributed.

Algorithmic fairness has most often been studied in one-shot problems. In applications such as lending or risk assessment, a decision is made once from data and fairness is a property of that single prediction. HITL adaptation is dynamic and sequential: an action taken now

reshapes future human states and future actions, preferences vary both between people and within a person over time, and a policy that is fair at one instant can become discriminatory later. Fairness is therefore defined here over time, as minimizing the discrepancy in people's accumulated experience, not their experience at any single step.

Two papers approach that goal differently. The first, **FinA** (*Fairness in Adverse Effect*) [Zhao and Elmalaki, 2024], is optimization-based. Drawing on prospect theory and the psychology of loss aversion (losses are felt far more acutely than equivalent gains), FinA defines an *adverse effect* that measures how strongly an action departs from a person's preference, and formalizes it in instantaneous, accumulated long-term, and perception-based forms. Across five formulations, FinA selects, at each state, the adaptation action that minimizes disparity in adverse effect across the people involved. The second, **FAIRO** [Zhao et al., 2024], is learning-based. Instead of re-solving an optimization at every step, FAIRO decomposes the fairness objective into temporally abstracted sub-goals using the Options reinforcement-learning framework: each option is responsible for raising whoever is currently least satisfied and terminates once that person is no longer the worst off, which yields a policy that repeatedly redirects attention toward whoever is falling behind. Both [Zhao and Elmalaki, 2024] and [Zhao et al., 2024] appeared at the ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS) 2024.

Section 2.2 discusses FinA and Section 2.3 discusses FAIRO. Section 2.4 discusses what the two establish jointly.

2.2 F*in*A: Fairness of Adverse Effects in Decision-Making of Human-Cyber-Physical Systems

2.2.1 Introduction

The ubiquitous integration of smart technologies into our daily lives offers unprecedented opportunities but also presents a lot of challenges. A key challenge lies in understanding the interplay between humans and Cyber-Physical Systems (CPS) to shape the societal consequences of future CPS technologies. As we move towards a future defined by the coexistence of humans and smart technologies, understanding their interactions is essential, as suggested by recent studies Annaswamy et al. [2023a,b]. Within the realm of Human-Cyber-Physical Systems (HCPS), one central challenge emerges when CPS decisions can affect individuals with diverse preferences and perceptions within the same environment. This scenario is common in systems like smart buildings, smart cities, smart traffic management, and smart crowd control, where fairness, privacy, equity, and personalization issues intersect, sparking new societal tensions Lee et al. [2017], Wang et al. [2020], Stangor [2014].

Existing research in sociology, particularly Social Exchange Theory Homans [1974], Cook and Emerson [1987] and Equity Theory Adams [1963] has substantiated a direct link between "equity" and prosocial behavior. The higher an individual perceives a system as equitable, the greater the likelihood of that individual engaging in prosocial behavior Thibaut and Kelley [1959], Redmond [2015], Adams [1963]. This, in turn, significantly impacts the overall performance of the system. Specifically, the greater the number of individuals who engage in prosocial behavior, the higher the likelihood of compliance and acceptance of the system's decisions, ultimately leading to an enhancement in overall system performance Cialdini [2007, 2016], Huijts et al. [2012]. Given that CPS decision-making often aims to optimize system performance while adhering to operational constraints, it is essential to develop formal metrics and objective functions that enhance the human perception of these decisions, ultimately

fostering more prosocial behavior and improving overall system performance.

In F*in*A, we are interested in formalizing some of the equity objectives in decision-making HCPS. We will start by exploring a notion of fairness from the equity perspective. In particular, we will focus on what we term **Fairness-in-Adverse-Effects (F*in*A)**. Decision-making agents in HCPS employ a range of control actions. However, these control actions can have different adverse effects on a diverse population, each with its own preferences. Hence, the HCPS needs to adapt its decision-making to continuously match different populations across time and ensure that it meets the preferences of as many populations as possible. Our motivation in examining the adverse effects or the harmful impact of decision-making in HCPS stems from the psychology concept **"loss aversion**–Losses loom larger than gains" which implies that losses can be twice as powerful, psychologically, as gains Kahneman and Tversky [2013]. This concept underscores the significance of minimizing adverse effects to promote fairness within HCPS.

2.2.2 Related Work

HCPS systems are centered on the challenge of designing adaptable, real-time decision-making processes that take into account the social context, including considerations of fairness, social welfare, ethical concerns, and societal norms Sztipanovits et al. [2019], Khargonekar and Sampath [2020]. A substantial body of work in the field of game theory explores various facets of fairness, often framed as incentive markets among competing entities or communities striving to achieve fairness Ratliff and Fiez [2020], Ratliff et al. [2018]. In the domain of machine learning, interventions to enhance fairness have been introduced, aiming to ensure that models' decisions are devoid of discrimination Friedler et al. [2019], Hashimoto et al. [2018], Chouldechova and Roth [2018], Kannan et al. [2018], Goel et al. [2018]. In the context of decision-making systems, where agents express favoritism for one action over another, questions surrounding fairness become even more significant, especially

within multi-agent systems Jabbari et al. [2017], Joseph et al. [2016], Yu et al. [2019], Gillen et al. [2018], Siddique et al. [2020], Shin et al. [2017], Jiang and Lu [2019], Hughes et al. [2018]. However, imposing fairness constraints as static, one-time decisions akin to conventional supervised learning methods while neglecting dynamic feedback and long-term consequences, particularly in sequential decision-making systems, can inadvertently lead to disparities that affect specific sub-populations Creager et al. [2020], Liu et al. [2018].

Recent research has also shed light on the long-term ramifications of Reinforcement Learning (RL), revealing that addressing control decisions' immediate effects in single steps does not ensure fairness in subsequent decision actions Kannan et al. [2019], Milli et al. [2019]. Nevertheless, a significant portion of this research has predominantly focused on fairness through the lens of equality, with an emphasis on eliminating favoritism or bias within the system, and relatively less attention has been given to the concept of fairness from an equity perspective, particularly in the context of sequential decision-making Mehrabi et al. [2021]. Notably, ensuring fairness in sequential decision-making systems becomes increasingly complex as policies deemed fair at one point may inadvertently become discriminatory over time due to shifts in human preferences influenced by the inter- and intra-human variation Elmalaki [2021].

The concept of "group fairness" has been introduced in the literature to tackle fairness concerns arising when the same adaptive model impacts multiple individuals. One notion is "Equalized Odds" which concentrates on achieving a level of uniform prediction accuracy across various groups, primarily within the context of binary classification tasks. The central objective is to ensure that a predictive model exhibits comparable true positive rates (sensitivity) and true negative rates (specificity) across diverse groups Hardt et al. [2016]. A second notion is "Equal opportunity,", which aims to guarantee that a predictive model affords an equal likelihood of beneficial outcomes for all groups. It places a specific requirement on the true positive rate, mandating that it should be approximately equivalent for

each group Hardt et al. [2016].

Prior research in CPS has explored fairness in various ways. For instance, fairness-aware resource allocation and scheduling algorithms have been developed for CPS, addressing issues like energy consumption and real-time constraints while considering equitable distribution among system components Iosup et al. [2017]. The concept of fairness in communication protocols for CPS has been established through strategies to ensure fair access to network resources for different devices and applications Huaizhou et al. [2013]. Furthermore, fairness challenges in decentralized CPS environments have been examined, focusing on ensuring fair decision-making in multi-agent systems Ho et al. [2020].

However, it's worth noting that much of the existing CPS literature concentrates on system-level performance and efficiency, often at the cost of individual-level fairness considerations. In contrast, F*in*A aims to delve deeper into the aspects of fairness within HCPS, addressing the interplay between human preferences, the temporal dimension of adverse effects, and perceptions of fairness. This approach allows for a more comprehensive understanding of fairness in the context of HCPS decision-making, particularly in systems where individuals' preferences and perceptions can evolve over time.

2.2.2.1 Contributions of F*in*A

The contributions of F*in*A can be summarized as follows:

- **Fairness-in-Adverse-Effect (F*in*A)**: We introduce a novel concept known as "Fairness-in-Adverse-Effect (F*in*A)." F*in*A takes a fresh perspective by considering equity in adverse effects within Human-Cyber-Physical Systems (HCPS) decision-making. We address scenarios where adaptive decisions made within HCPS can impact multiple individuals with diverse preferences. By formalizing F*in*A, we provide a means to ensure that the adverse effects of decision-making are distributed fairly among the system's users
- **Long-term effects:** Our work extends beyond the immediate effect of CPS decision-

making. We delve into the temporal dimension of adverse effects, recognizing that the impact of these decisions can have lasting consequences. The interplay between human preferences, historical data, and the evolving perception of fairness adds complexity to the notion of fairness. We introduce five different approaches to formalize F*in*A within CPS decision-making to account for the relationship between the instantaneous adverse effects, long-term adverse effects, and the overall perception of fairness.

- **Generalization to different HCPS setups:** We acknowledge that the nature of adverse effects and fairness considerations can vary across different domains. Therefore, we offer a general formalization of F*in*A that can be applied to various HCPS scenarios. Additionally, we conduct thorough evaluations in the domain of smart home to illustrate the trade-offs between various interpretations and implementations of F*in*A. This demonstrates the flexibility and effectiveness of our approach across diverse HCPS settings.

The remainder of this section is organized as follows: We introduce the notion of Fairness-in-Adverse-Effects (F*in*A) within Human-Cyber-Physical Systems (HCPS) in Section 2.2.3. We formally define F*in*A using five different approaches in considering the instantaneous and the long-term adverse effects while considering the human perception of fairness in Sections 2.2.3.2, 2.2.3.3, 2.2.3.4, 2.2.3.5 and 2.2.3.6. Afterward, we numerically evaluate these approaches using a smart home HCPS application in Section 2.2.4.

2.2.3 Approach

We consider a CPS depicted in Figure 2.1, which serves multiple individuals sharing the same CPS environment, each with different preferences. The control action generated by the decision-making agent in CPS can cause different adverse effects on those individuals.

To achieve Fairness-in-Adverse-Effects (F*in*A) within Human-Cyber-Physical Systems (HCPS), we propose five distinct approaches to guide CPS decision-making when selecting control actions affecting multiple humans sharing the same environment. These approaches are

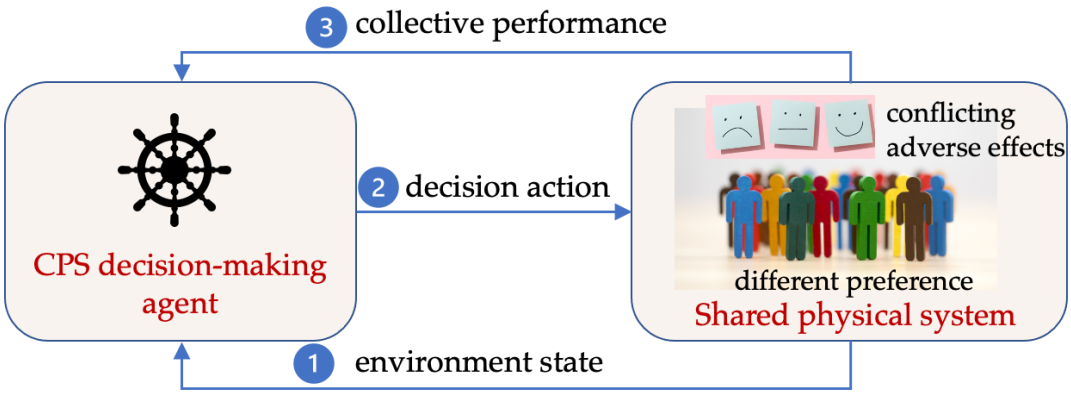


Figure 2.1: Human-Cyber-Physical-System with multiple individuals sharing the same environments with different preferences. The decision-making agent control action can cause different adverse effects on those individuals.

rooted in the recognition that individuals who exhibit pro-social attitudes might be willing to tolerate certain discrepancies between their preferences and CPS actions for a limited duration Thibaut and Kelley [1959]. However, it is important to acknowledge that this willingness to forgive may not be indefinite, as the magnitude and persistence of discrepancies play a pivotal role in shaping individuals' perception of fairness Thibaut and Kelley [1959]. Moreover, the extent of an individual's willingness to forgive is contingent upon the CPS's responsiveness in addressing and rectifying such discrepancies Cook and Emerson [1987].

Our first approach formalizes F*in*A by examining the concept of instantaneous adverse effects, focusing on the immediate impact of CPS control actions (Section 2.2.3.2). In this context, we recognize that individuals may exhibit a degree of patience when faced with minor or unintentional discrepancies between their preferences and CPS actions. This approach is predicated on the assumption that in scenarios of minimal and short-term discrepancies, individuals may still perceive the CPS as acting in their best interest.

Nonetheless, the second approach acknowledges that as the severity and persistence of discrepancies increase, individuals, even those with pro-social attitudes, may become less forgiving (Section 2.2.3.3). Thus, the historical aspect of adverse effects becomes a critical factor.

It is during extended periods of inconsistency that individuals may experience a diminishing willingness to tolerate disparities. This approach takes into account the temporal dimension of adverse effects, recognizing that extended discrepancies may erode the perception of fairness Jain et al. [1984].

In our third approach, we examine a balance between the instantaneous adverse effect and the historical record of adverse effects (Section 2.2.3.4). By combining these two dimensions, we aim to provide a tradeoff that accounts for both short-term variations and long-term consequences. This approach is particularly valuable in situations where CPS must navigate the delicate balance between immediate and lasting impact.

Building on the well-established literature on fair resource distribution, we acknowledge that fairness is not synonymous with equal resource allocation Jain et al. [1984]. Humans' perception of fairness is intrinsically tied to how they compare their resource distribution with that of others in the same system. In the fourth approach, we consider the collective perception of fairness among individuals who share the same CPS environment as a metric for formalizing FinA (Section 2.2.3.5). This approach recognizes that individuals may be more forgiving of discrepancies if they perceive that others are experiencing similar variations in resource allocation.

Lastly, our fifth approach introduces a tradeoff between human perception of fairness and a budget of allowable discrepancy between individual preferences and the applied CPS control action (Section 2.2.3.6). By imposing limits on the magnitude of permissible discrepancies, this approach provides a tradeoff between accommodating individual preferences and ensuring collective fairness.

2.2.3.1 Fairness-in-Adverse-Effect (FinA) Setup

In all of these five approaches, we consider a society S that consists of N different individuals and a CPS that serves this society by providing shared control actions a that may be tailored

toward the preferences and behavior of some of those individuals. We assume that every individual $n \in N$ has a different set of g preferred actions $A_n = \{a_1^n, a_2^n, \ldots, a_g^n\}$ that can serve them better[1]. Suppose an adverse effect $v_n(a)$ on individual n occurs due to the chosen control action a. An initial mechanism to measure the **adverse effect** on each $v_n(a)$ is to assume that a preferred action $a_g^n \in A_n$ is inversely proportional to its adverse effects. That is, to measure the adverse effects of a chosen control action $v_n(a)$ on human n, we can use the distance between the set of preferred actions $\mathcal{A}_n$ and the chosen control action a, i.e.,

$$v_n(a) = \max_{a_g^n \in \mathcal{A}_n} \|a_g^n - a\|. \tag{2.1}$$

2.2.3.2 Approach I: Formalizing the definition of FinA with instantaneous effect

In our first approach, we formalize FinA by examining instantaneous adverse effects, delving into the immediate consequences of CPS control actions.

Hence, an initial definition of FinA can be:

$$\text{F}in\text{A} = \min_{a \in \mathcal{A}} \|\mathbf{v}(a) - \frac{1}{N}\mathbb{1}^T\mathbf{v}(a) \otimes \mathbb{1}\| + \lambda\|\frac{1}{N}\mathbb{1}^T\mathbf{v}(a)\|, \tag{2.2}$$

where $\mathbf{v}(a) = [v_1(a), v_2(a), \ldots, v_N(a)]^\intercal$ and $\mathcal{A} = \bigcup_{n=1}^{N} A_n$. In other words, Equation (2.2) aims to choose the control action a, out of all possible A_n, that minimizes the difference between the individual adverse effects on every individual v_n and the average of the adverse effects across all individuals $\frac{1}{N}\sum_{n=1}^{N} v_n(a) = \frac{1}{N}\mathbb{1}^T\mathbf{v}(a)$. Indeed, one trivial solution will be to increase the adverse effect on all individuals to achieve the same average. Therefore, the second term in Equation (2.2) asks that the chosen control action also aims to minimize the

[1]While in our first definition A_n we assume a discrete set of preferred actions, this can be extended to a continuous range of preferred actions.

average adverse effect [2].

2.2.3.3 Approach II: Formalizing the definition of F*in*A using long-term temporal variations in adverse effect

In the second approach, the historical context of adverse effects plays a pivotal role. Prolonged instances of inconsistency are where individuals may exhibit a reduced willingness to endure disparities. This approach places a significant emphasis on the temporal dimension of adverse effects, acknowledging that extended discrepancies can undermine the perception of fairness Jain et al. [1984].

We define a long-term adverse effect $\mathbf{v}_n$ by monitoring the adverse effect for every applied action a over a time horizon T on human n as follows:

$$\mathbf{v}_n = [v_n^0, v_n^1, \ldots, v_n^{T-1}]^\intercal, \tag{2.3}$$

where v_n^j represents the adverse effect occurred at time j for human n. However, to focus more on the recent adverse effects, we assign different weights to v_n^j and calculate an accumulated value u_n^t that represents the current history of adverse effects on human n from time $t-T$ till $t-1$. Accordingly, by assigning the weight $\frac{j}{T}$ to v_n^j, the most recent adverse effects have more contribution to the the accumulated value u_n^t.

$$u_n^t = \frac{1}{T}\sum_{j=0}^{T-1} \frac{j}{T} v_n^j, \text{ for } n = 1, 2, \ldots, N \tag{2.4}$$

[2]We used L^2 norm for all $\| \bullet \|$ notations.

The historical adverse effect on all N individuals can be represented by:

$$\mathbf{u} = [u_1^t, u_2^t, \ldots, u_N^t]^\intercal \tag{2.5}$$

We can then formally define FinA as:

$$\begin{aligned} \text{F}in\text{A} = & \min_{a \in \mathcal{A}} \ \mathcal{B} \\ & s.t. \ \mathbf{v}(a) < \mathcal{B} - \mathbf{u} \end{aligned} \tag{2.6}$$

In this equation, $\mathbf{v}(a) = [v_1(a), v_2(a), \ldots, v_N(a)]^\intercal$ represents the current adverse effect of action a on each individual. To elaborate, Equation (2.6) contains N constraints, with each constraint governing the current adverse effect $v_n(a)$ to remain below a specific budget $\mathcal{B}$. The intuition here is to use this budget $\mathcal{B}$ as the total amount of adverse effect in the past T history. However, it's essential to note that this budget $\mathcal{B}$ is gauged on the historical adverse effects of each individual, denoted as u_n^t. In this context, FinA tries to identify the minimum budget $\mathcal{B}$ that satisfies all N constraints. Consequently, if a human individual, say n, has a substantial historical adverse effect value, the constraint $\mathcal{B} - u_n^t$ will direct the optimization process toward finding an action a that results in a minimal $v_n(a)$. This approach is designed to steer the optimization's focus towards individuals with higher historical adverse effect values, encouraging the selection of new actions that minimize their adverse effects.

2.2.3.4 Approach III: Formalizing the definition of FinA as a tradeoff between instantaneous and long-term adverse effect

In our third approach, we delve into the balance between immediate adverse effects and cumulative historical adverse effects. To achieve this, we augment Equation (2.2) with a term that considers the historical adverse effects, denoted as u_n^t as defined in Equation (2.4).

The extended formulation can be expressed as:

$$\begin{aligned} \mathrm{F}in\mathrm{A} = \min_{a\in\mathcal{A},\mathbf{b}} \; & \alpha\Big(\|\mathbf{v}(a) - \frac{1}{G}\mathbb{1}^T\mathbf{v}(a)\otimes\mathbb{1}\| + \lambda\|\frac{1}{G}\mathbb{1}^T\mathbf{v}(a)\|\Big) \\ & + (1-\alpha)\mathbf{u}^\intercal\mathbf{b}, \\ s.t. \; & \mathbf{v}(a) < \mathbf{b} \end{aligned} \tag{2.7}$$

In this formulation, $\mathbf{b} = [b_1, b_2, \ldots, b_N]^\intercal$ and $\mathbf{u} = [u_1^t, u_2^t, \ldots, u_N^t]^\intercal$. Essentially, the first term in Equation (2.7) is from the instantaneous adverse effect (Equation (2.2)), and then we introduce N new optimization variables, b_n, for each individual n, which represents the budget allocated for the adverse effect ($v_n(a)$) for each individual n. Importantly, these budgets $\mathbf{b}$ are weighted by the values of the long-term adverse effects $\mathbf{u}$.

In other words, we have N constraints for different budgets for adverse effects that are bounded for every individual ($\mathbf{v}(a) < \mathbf{b}$). Accordingly, F$in$A needs to minimize these budgets $\mathbf{b}$ to minimize the overall adverse effects. However, the upper bound for these budgets is weighted by the accumulated historical adverse effects. Hence, the term $\mathbf{u}^\intercal\mathbf{b}$ is appended to the definition of FinA. To modulate the trade-off between the immediate and the long-term adverse effects, we introduce parameter α where $\alpha \in [0, 1]$.

2.2.3.5 Approach IV: Formalization the definition of FinA using human perception of fairness

Drawing upon extensive research in the realm of equitable resource distribution, we recognize that fairness does not necessarily mean equal resource allocation Jain et al. [1984]. Instead, the perception of fairness in individuals is fundamentally linked to how they gauge their own resource distribution concerning that of their peers within the same system. In particular, fairness in this setup can be viewed generally as "variances" Jain et al. [1984] of the "utility" shared by individuals. Hence, we exploit the notion of the coefficient of variation (CoV) of

the utilities Jain et al. [1984]. In our setup, we define this utility for human n at time t as the temporal accumulated adverse effect caused by the control actions of the CPS in the shared environment, denoted by u_n^t as expressed in Equation (2.4).

$$CoV_{\mathbf{u}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\frac{(u_n - \bar{\mathbf{u}})^2}{\bar{\mathbf{u}}^2}} \tag{2.8}$$

In Equation (2.8), $\bar{\mathbf{u}} = \frac{1}{N}\sum_{n=1}^{N} u_n^t$ represents the average utility (accumulated adverse effect at time t) of all N humans. The system is said to be more fair if and only if the CoV is smaller. The value of CoV can be anywhere between 0 and infinity. Hence, we use the fairness index (FI) transformation to have a value between 0 and 1 to be easily interpreted as a fairness percentage. In other words, if FI is 1, it means the system is 100% fair, otherwise, if disparity increases between individuals, this FI value will decrease Jain et al. [1984].

$$FI_{\mathbf{u}} = \frac{1}{1 + CoV_{\mathbf{u}}^2} \tag{2.9}$$

Accordingly, we can define FinA to maximize the fairness index. The core idea is to minimize the reciprocal of this fairness index, represented as y in the optimization.

$$
\begin{aligned}
\mathrm{F}\mathit{in}\mathrm{A} &= \min_{a \in \mathcal{A}} \; y \\
s.t.\; y &\geq 1 + \frac{1}{N} \sum_{n=1}^{N} \left(\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \right)^2 \\
& \frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \leq \epsilon, \text{ for } n = 1, 2, \ldots, N.
\end{aligned} \tag{2.10}
$$

where:

$$u_n = u_n^t + v(a)$$

In this formulation, u_n^t is a constant value that represents the accumulated history of the adverse effect up till time $t-1$ (Equation (2.4)) before applying the new a that will add the new adverse effect $v(a)$ on human n. We also add the constraint $\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \leq \epsilon$, for $n = 1, 2, \ldots, N$ that sets a limit on how much each individual's adverse effect can deviate from the average adverse effect. The parameter ϵ defines the maximum allowable difference.

2.2.3.6 Approach V: Formalizing the definition of FinA using the human perception of fairness with a tradeoff for a budget for long-term adverse effect

Lastly, our fifth approach introduces a tradeoff between human perception of fairness and a budget of allowable discrepancy between individual preferences and the applied CPS control action. In particular, we combine our definition of FinA in Equation (2.6) and Equation (2.10) to provide a tradeoff between fairness index and adverse effect budget $\mathcal{B}$.

$$
\begin{aligned}
&\mathrm{F}in\mathrm{A} = \min_{a \in \mathcal{A}} \ \alpha.y + \beta.\mathcal{B} \\
&s.t.\ y \geq 1 + \frac{1}{N} \sum_{n=1}^{N} \left(\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \right)^2 \\
&\mathbf{v}(a) < \mathcal{B} - \mathbf{u}
\end{aligned}
\tag{2.11}
$$

where:

$$u_n = u_n^t + v(a)$$

The α and β weights allow us to adjust the tradeoff between fairness index and adverse effect budget.

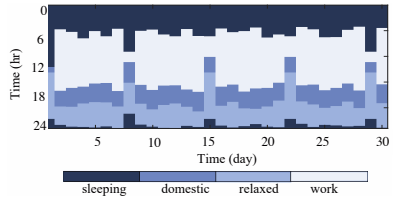


Figure 2.2: Human 1 activity pattern.

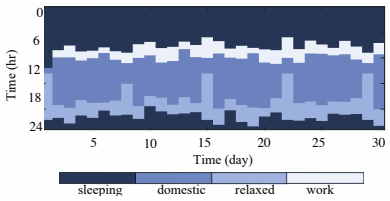


Figure 2.3: Human 2 activity pattern.

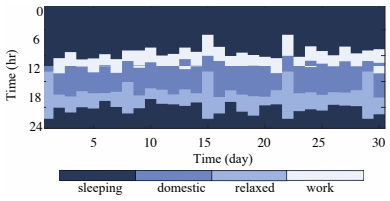


Figure 2.4: Human 3 activity pattern.

2.2.4 Evaluation

We designed an HCPS application in the domain of smart house. Recent literature focuses on enhancing human satisfaction in smart heating, ventilation, and air conditioning (HVAC) systems by employing various techniques to adjust the set-point based on human activity and preferences Jung and Jazizadeh [2017], Elmalaki [2021]. These HCPS systems consider the current state and individual preferences, such as body temperature changes during sleep or physical activity. To evaluate different approaches of FinA in this application, we consider a setup where multiple humans share a house with a single HVAC system, and their activities determine individual HVAC set-point preferences. Humans can be in the same room or

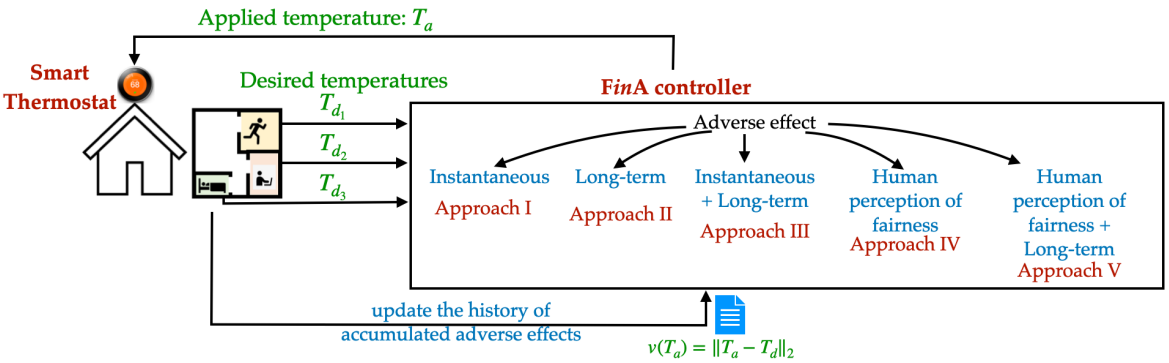


Figure 2.5: A smart house with three humans. Each human has a different activity, which requires different desired indoor temperature setpoints T_d. An external controller running FinA selects the applied setpoint T_a based on the calculations of instantaneous and long-term adverse effects.

separated rooms as long as they are in the same shared application space that is controlled by the same HVAC.

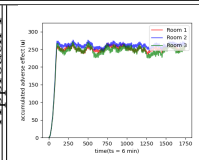
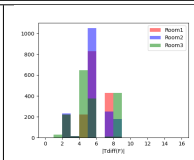
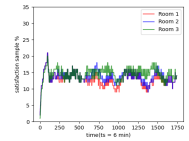
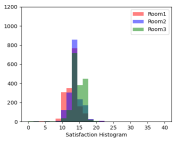
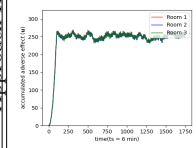
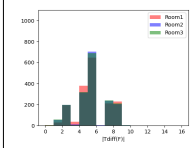
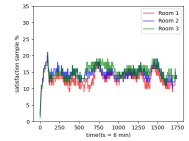
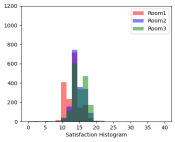
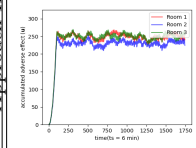
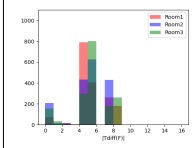
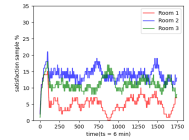
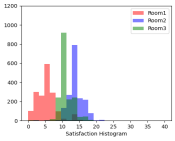
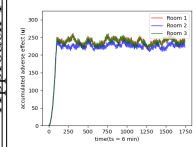
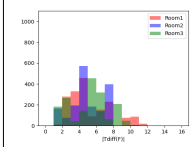
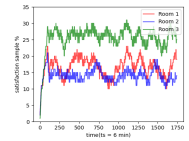
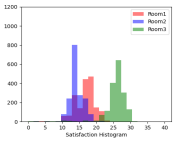
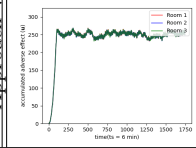
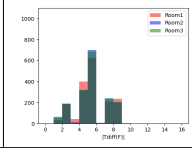
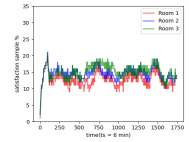
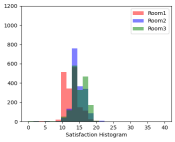
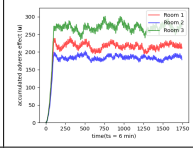
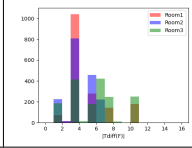
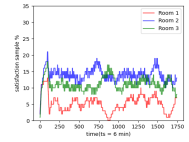
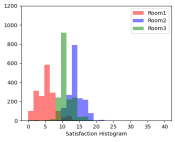
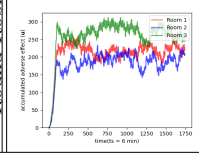
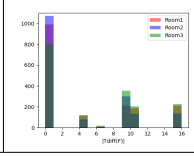
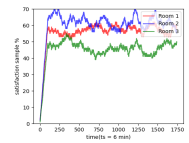
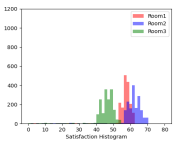
We exploited recent work in the literature Elmalaki [2021] that simulated a thermodynamic model of a house incorporating the house's shape and insulation type. To regulate indoor temperature, a heater and a cooler with specific flow temperatures ($50°c$ and $10°c$) were employed. A thermostat maintained the indoor temperature within $2.5°c$ around the desired set point. An external controller controls the setpoint running the optimization of FinA. A pictorial figure of the application setup is shown in Figure 2.5.

We implemented our proposed five approaches using CVXPY, a Python-embedded modeling language for convex optimization problems Diamond and Boyd [2016].

The human was modeled as a heat source, with heat flow dependent on the average exhale breath temperature (EBT) and the respiratory minute volume (RMV). These parameters depend on human activity Carroll [2007]. We simulated three humans with four activities: sleeping, relaxing, medium domestic work, and working from home. Randomness was introduced by allowing multiple activity choices during the same time slot. The different activity schedules depicted in Figures 2.2, 2.3, and 2.4. The humans were simulated in separate

rooms as seen in Figure 2.5, each exhibiting unique behavioral patterns: (1) h_1 followed an organized and repetitive weekly routine, (2) h_3 had a more random and unpredictable life pattern, and (3) h_2 displayed intermediate randomness, alternating between sleeping, working from home, domestic activities, and relaxation. The Mathworks thermal house model was extended to include a cooling system and a human model[3].

Table 2.1: Comparison between all the different five approaches of F*in*A, Mean approach, and Round Robin.

	Accumulated adverse effect($\mathbf{u}$)	Hist. temperature difference $\lvert T_{diff} \rvert$	Satisfaction rate (SR) %	Hist. satisfaction rate (SR)
Approach I				
Approach II				
Approach III				
Approach IV				
Approach V				
Mean				
Round Robin				

[3]While more complex simulators like EnergyPlus Gerber [2014] exist, considering energy consumption and electric loads, we opted for a simpler model to assess F*in*A.

The desired preferred action (temperature setpoint) per human $a_g^n = T_n$, for $n = 1, 2$, and 3, can be obtained through fixed policy configuration. We exploit existing approaches Taherisadr et al. [2023] for estimating the desired HVAC setpoint based on activity and thermal comfort. The desired setpoints for the considered activities are domestic activity (72°F), relaxed activity (77°F), sleeping (62°F), and work from home (67°F). These setpoints aim to enhance thermal comfort Fanger [1970]. The shared control action space (applied temperature setpoint) is all possible temperatures ranging from the minimum preferred action to the maximum preferred action as defined in Section 2.2.3.1.

2.2.4.1 Experiment setup

In this application, we used the difference between the desired temperature (T_d) and the applied temperature (T_a) as a measure of the adverse effect:

$$v(a) = \|T_a - T_d\|_2, \text{ where } T_a \in [60 - 80]^\circ F$$

We use the most recent 100 samples for our history of adverse effects for the three humans $\mathbf{v}_1, \mathbf{v}_1$ and $\mathbf{v}_3$ (as explained in Equation (2.3)) with sampling time $t_s = 6$ min. Hence, every t_s, we compute $\mathbf{u} = [u_1, u_2, u_3]^\intercal$ for the three humans (as explained in Equation (2.5)), where $u_n = \frac{1}{100} \sum_{j=0}^{99} \frac{j}{100} v_n^j$, and $v_n^j = \|T_a^j - T_{d_1}^j\|_2$ for $n = 1, 2$, and 3.

The simulation was executed using $3,000$ samples, roughly equivalent to approximately 12 days. This extended duration allowed us to accurately capture changes in the behavioral patterns of the three individuals. At every time step, we check if the desired temperatures of the three humans are not identical then we run the optimization solver for FinA then update all the corresponding u.

We consider that the human is satisfied if the T_a is within $2.5^\circ F$ difference from the desired temperature. We measure the satisfaction rate by considering the 100 sample window in $\mathbf{v}_n$.

Hence, the satisfaction rate (SR) for human n is computed every t_s computed as:

$$SR_n = \sum_{j=0}^{99} \mathbb{1}(v(a) \leq 2.5), \text{ for all } v(a) \in \mathbf{v}_n$$

Hence, the SR can give us a measure in % since the total number of samples we consider is 100. We use this SR to compute CoV and FI as a function of the SR similar to Equations (2.8) and (2.9) respectively.

$$CoV_{\mathbf{SR}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\frac{(SR_n - \overline{\mathbf{SR}})^2}{\overline{\mathbf{SR}}^2}}, \tag{2.12}$$

where $N = 3$ and $\mathbf{SR} = [SR_1, SR_2, SR_3]^\intercal$ computed for the three humans every t_s. Similarly, we consider the $FI_{\mathbf{SR}}$ as follows:

$$FI_{\mathbf{SR}} = \frac{1}{1 + CoV_{\mathbf{SR}}^2} \tag{2.13}$$

Furthermore, we compared the five proposed approaches for FinA with two more approaches:

- **Mean approach:** In this case, the applied temperature T_a is the mean of the desired temperature from the three humans.
- **Round Robin**: In this case, the applied temperature T_a is selected in rotation between the desired temperature from the three humans.
- **FaiRIoT Elmalaki [2021]**: We compare with the state-of-the-art FaiRIoT, which uses hierarchical reinforcement learning to assign weights to the desired actions to compute the applied action. Hence, $T_a = \sum_{n=1}^{3} w_n T_{d_n}$.

In all experiments, we set the trade-off parameter $\alpha = 0.5$, as defined in Equation 2.7.

Table 2.2: Comparison of the overlap area percentage, Satisfaction JSD , and average Fairness Index (FI) and the average coefficient of variation (CoV) of adverse effect($\mathbf{u}$) and satisfaction($\mathbf{SR}$), respectively.

	$\lvert T_{diff}\rvert$ overlap%	SR JSD	$Avg.$ $FI_{\mathbf{u}}$	$Avg.$ $CoV_{\mathbf{u}}$	$Avg.$ $FI_{\mathbf{SR}}$	$Avg.$ $CoV_{\mathbf{SR}}$
Appr. I	22.4%	0.086	0.998	0.026	**0.994**	**0.057**
Appr. II	**86.5%**	**0.010**	**0.999**	**0.004**	**0.994**	0.066
Appr. III	37.6%	0.639	0.998	0.038	0.870	0.365
Appr. IV	19.2%	0.659	0.998	0.027	0.929	0.624
Appr. V	83.4%	0.139	**0.999**	**0.004**	0.992	0.077
Mean	24.8%	0.648	0.974	0.157	0.868	0.370
RR	68.4%	0.723	0.973	0.160	0.984	0.124

2.2.4.2 Results

We plot in Table 2.1, the accumulated adverse effect ($\mathbf{u} = [u_1, u_2, u_3]^\intercal$), the histogram of the absolute temperature difference between $|T_{diff}| = |T_a - T_d|$, the satisfaction rate (SR), and the histogram of the satisfaction rate (SR), across all approaches for the three individuals in three rooms. First column in Table 2.1 compares the differences in the individual's adverse effect ($\mathbf{u} = [u_1, u_2, u_3]^\intercal$) across all approaches. Approach II and V show the smallest difference which is also reflected in average $CoV_{\mathbf{u}}$ in Table 2.2. Approach IV has a higher $CoV_{\mathbf{u}}$ (0.027) but it can bound $\mathbf{u}$ within a smaller value compared with other approaches observed in Table 2.1.

We compare the distribution of $|T_{diff}|$ across all the approaches in Table 2.1 second column. Approach II has the highest overlap percentage 86.5% as calculated in Table 2.2 which indicates that this approach can make all 3 rooms have a more similar experience compared with other approaches. Round robin (RR) has a large overlap percentage due to the fact that each room can have a $|T_{diff}| = 0$ on its turn in the round. However, RR will result in

significant $|T_{diff}|$, which is larger than $10°F$ in a notable number of the samples.

Table 2.1 third and fourth columns present the satisfaction rate (**SR**) across all approaches. We report the Jensen-Shannon Divergence (JSD) of the histogram for **SR** in Table 2.2. Approach II has the lowest JSD, indicating closer SR across rooms. RR has the highest overall SR but it has the highest JSD indicating no fairness in the SR among 3 rooms[4].

We summarize the average values of FI and CoV for long-term adverse effect u and satisfaction rate SR in Table 2.2 with more detailed results shown in Table 2.3. The fairness index (FI) is a metric ranging from 0 to 1, where 1 means absolute fair as explained in Section 2.2.3.5. In particular, as shown in Table 2.3, Approach I, II, III, and IV have $FI_{\mathbf{u}}$ values close to 1 and their $CoV_{\mathbf{u}}$ values are less than 0.04. On the contrary, Mean and RR have $FI_{\mathbf{u}}$ around 0.97 and a $CoV_{\mathbf{u}}$ of 0.16.

2.2.4.3 Comparison between these approaches

Approach II and V have the best $FI_{\mathbf{u}}$ and $CoV_{\mathbf{u}}$, which indicates better fairness in terms of long-term adverse effect u. Approach III and IV have comparable $FI_{\mathbf{u}}$ and $CoV_{\mathbf{u}}$ but their fairness in terms of satisfaction rate (**SR**), measured in metrics $FI_{\mathbf{SR}}$, and $CoV_{\mathbf{SR}}$, is not improved. Based on these analysis from Tables 2.1, 2.2, and 2.3, we observe that Approach II, and Approach V provide the best results in terms of $FI_{\mathbf{u}}$, and $CoV_{\mathbf{u}}$, while Approach I provides better results in terms of $FI_{\mathbf{SR}}$, and $CoV_{\mathbf{SR}}$.

2.2.4.4 Implementation and Execution time

We used M1 Pro chip as the platform to run our optimization solver. We use open source solver ECOS in CVXPY Diamond and Boyd [2016]. Since we only call the optimization solver when the desired actions between the N humans are different, we measure the execution

[4]The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. The JSD is symmetric and always non-negative, with a value of 0 indicating that the two distributions are identical, and a value greater than 0 indicating that the two distributions are different.

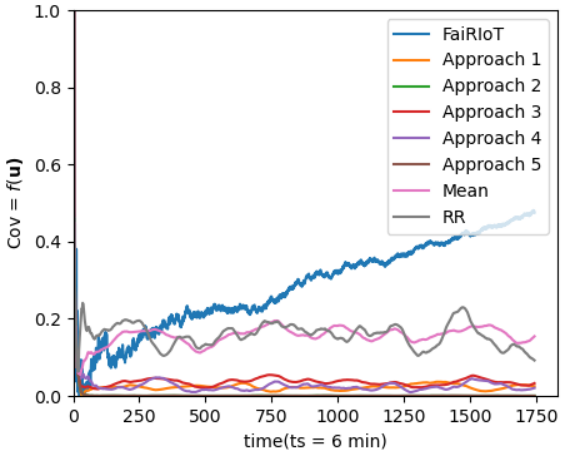


Figure 2.6: Coefficient of Variation (CoV) comparison between FaiRIoT, Mean, Round Robin(RR), and all approaches of FinA .

time only when the solver is called. Then, we averaged 1000 calls for the solver for different FinA approaches. The average execution time per 1000 calls for the optimization solver for Approach I, II, III, IV and V are 6.308s, 8.249s, 9.876s, 10.355s and 10.217s respectively. This indicates that around 1ms overhead on average to call the solver. We used `time()` function in Python 3 to measure the execution time.

2.2.4.5 Compare with the state of the art FaiRIoT Elmalaki [2021]

The closest to our approach is FaiRIoT which computes the applied action through a weighted sum of all the desired actions by the N individuals $T_a = \sum_{n=1}^{N} w_n T_{d_n}$. FaiRIoT uses a notion of utility which is the average weight assigned by a layer called "Mediator RL" for a particular human h over a time horizon $[0 : t]$. In particular, FaiRIoT measures the fairness of the Mediator RL using the coefficient of variation (CoV) of the human utilities. The Mediator RL is said to be more fair if and only if the CoV is smaller. Accordingly, in Figure 2.6, we compare the CoV in FaiRIoT with the $CoV_{\mathbf{u}}$ in all approaches in FinA. Approaches I - V achieve average CoV around 0.20, while FaiRIoT CoV is larger than 0.6. Approach II and IV has the lowest CoV at 0.04. **Hence, using FinA approaches improves the fairness where CoV is reduced by 66.7% on average.**

2.2.5 Discussion

In F*in*A, we proposed an initial mechanism that we used to define adverse effect $v_n(a)$ of action a on human n as described in Equation 2.1. This assumes that the human preferred action is inversely proportional to its adverse effects. This definition can be gauged by the human perception of adverse effects. In particular, cognitive psychology offers insights on human perception. For example, Bounded Rationality Theory suggests that individuals satisfice rather than optimize decision-making Simon [1997]. Satisficing means seeking solutions that are "good enough" or satisfactory for a given situation rather than exhaustively exploring all possible options to identify the optimal choice. Hence, the adverse effect can be a time-varying function based on human perception of satisfaction.

2.2.6 Conclusion

Addressing fairness in decision-making not only aligns with the principles of ethical AI and responsible technology, but also highlights the importance of socially-aware CPS, as individuals are more likely to cooperate with, and ultimately accept, systems that they perceive to treat them fairly. Our approaches to formalizing F*in*A within CPS decision-making capture the interplay between human preferences, the temporal dimension of adverse effects, and perceptions of fairness. Recognizing the complexities of these interactions is essential for designing more equitable Human-Cyber-Physical Systems. These approaches offer a multifaceted perspective on addressing the challenges posed by the impact of CPS control actions on diverse individuals within shared environments.

2.3 FAIRO: Fairness-aware Sequential Decision Making for Human-in-the-Loop CPS

2.3.1 Introduction

The emerging technologies of sensor networks and mobile computing give the promise of monitoring the humans' states and their interactions with the surroundings and have made it possible to envision the emergence of human-centered design of cyber-physical systems (CPS) applications in various domains. This tight coupling between human behavior and computing enables a radical change in human life. By continuously developing a cognition about the environment and the human state and adapting/controlling the environment accordingly, a new paradigm for CPS systems provides the user with a personalized experience, commonly named Human-in-the-Loop (HITL) systems. With the increasing number of HITL CPS applications being controlled by artificial intelligence (AI) algorithms, the algorithmic fairness of such decision-making algorithms has drawn considerable attention in the last few years Kleinberg et al. [2018]. Nevertheless, the unique nature of HITL CPS opens a new frontier of algorithmic fairness issues that must be carefully addressed before the wide use of such technologies. In particular, the immense challenge in designing the future HITL CPS lies in respecting human rights and values, ensuring ethics and fairness, and meeting regulatory guidelines while safeguarding our environment and natural resources Annaswamy et al. [2023a].

The following summarizes the key distinctions between the existing literature on algorithmic fairness and the nature of Human-in-the-Loop (HITL) systems:

- **Fairness in static/singular decision-making vs fairness in dynamic/sequential decision making:** The current literature on algorithmic fairness primarily addresses the unfairness arising from biases in data and algorithms used in static systems, often employing supervised learning methods. A canonical example comes from a tool used

by courts in the United States to make pretrial detention and release decisions (COMPAS) Northpointe [2015]. Other applications include loan applications, employment processes and markets Mukerjee et al. [2002]. In contrast, HITL systems are dynamic, where actions taken at one time have consequences for future states and actions. Therefore, ensuring fairness in HITL systems requires considering the impact of decisions over time, leading to a sequential decision making problem. Neglecting the dynamic feedback and long-term effects in such systems, as commonly done in static decision-making, can harm sub-populations Creager et al. [2020].

- **Fairness in decisions (or equality) vs fairness in the impact of decisions (or equity):** Existing fairness definitions predominantly focus on equality, aiming to eliminate prejudice or favoritism based on individuals' characteristics. However, insufficient attention has been given to equity, which entails allocating resources to individuals or groups to support their success Mehrabi et al. [2021]. Equity becomes crucial in HITL systems. Hence, a shift from fairness defined in terms of equality to fairness based on equity is essential.

Motivated by these observations, FAIRO revisits fairness literature and emphasizes the importance of fairness in sequential decision making from an equity perspective for HITL systems. The main objective is to operationalize equity in the context of sequential decision making to develop improved adaptation algorithms tailored to HITL applications.

This section introduces **FAIRO**, a novel fairness-aware adaptation framework for sequential decision making designed for HITL systems. The framework specifically tackles the issue of fairness in situations where multiple humans share the same application space and are collectively impacted by adaptation decisions. While usually, these decisions aim to optimize overall system performance, they may inadvertently lead to undesired consequences as humans interact with the system or as the system's physical dynamics evolve.

2.3.2 Related Work

2.3.2.1 Fairness in decision-making systems

At the heart of HITL systems is achieving the objective of designing scalable, real-time decision-making mechanisms that are aware of the social context, such as the perceived notion of fairness, social welfare, ethics, and social norms Khargonekar and Sampath [2020]. A vast work in the game theory literature studies various notions of fairness between communities by defining incentive markets between competitors to achieve fairness Ratliff and Fiez [2020]. Fairness-enhancing interventions have been introduced to machine learning to ensure non-discriminatory decisions by the trained models Friedler et al. [2019], Hashimoto et al. [2018]. In particular, the question of fairness in decision-making systems where the agent prefers one action over another Zhao et al. [2023b], Elmalaki [2021] becomes more significant in multi-agent systems Jiang and Lu [2019]. However, imposing fairness constraints as a static, singular decision (as standard supervised learning methods do) while ignoring subsequent dynamic feedback or its long-term effect, especially in sequential decision making systems, can harm sub-populations Creager et al. [2020]. Recent work investigates the long-term effects of Reinforcement Learning (RL). It shows that modeling the instantaneous effect of control decisions for single-step bias prevention does not guarantee fairness in later downstream decision actions Kannan et al. [2019]. Unfortunately, all of this work focused on fairness from the lens of equality—where the target is to ensure no favoritism or bias is present in the system—with very little work that focused on fairness from the lens of equity (mostly in singular/static decision making as opposed to sequential decision making) Mehrabi et al. [2021]. Indeed, achieving fairness in sequential decision making systems becomes more complex since policies deemed fair at one point may become discriminatory over time due to variations in human preferences resulting from inter- and intra-human factors Elmalaki [2021]. FAIRO focuses on answering this question, especially for HITL systems.

2.3.2.2 Different notions of group fairness

The notion of "group fairness" is used in the literature to address the fairness problem when multiple humans are affected by the same adaptation model. While there are several definitions and approaches to defining group fairness, it's important to note that these approaches may have nuanced variations and can be interpreted differently depending on the context and specific application domain. We only summarize two widely used notions of group fairness: (1) equalized odds, which focuses on achieving similar prediction accuracy across different groups while considering binary classification tasks. It ensures that the true positive rate (sensitivity) and true negative rate (specificity) of a predictive model are comparable across different groups Hardt et al. [2016], and (2) equal opportunity: aims to ensure that the predictive model provides an equal chance of benefiting from positive outcomes for all groups. In particular, equal opportunity requires that the true positive rate for each group should be approximately equal Hardt et al. [2016]. While these two definitions primarily focus on binary classification tasks, FAIRO will exploit some of their ideas towards sequential decision-making and not specifically for classification tasks.

2.3.2.3 Multi-agent RL and hierarchical RL

Reinforcement Learning (RL) is a widely used approach for monitoring and adapting to human intentions and responses in various contexts Sadigh et al. [2017]. To account for individual variability and response times, approaches like multisample RL has been proposed Elmalaki et al. [2018]. Hierarchical reinforcement learning (HRL) decomposes complex learning tasks into manageable components by using a hierarchical structure. The high-level policy selects optimal sub-tasks, considered high-level actions, while the lower-level policy focuses on solving these sub-tasks using reinforcement learning techniques. This decomposition strategy transforms long timescale tasks into multiple shorter timescale sub-tasks, potentially simplifying individual sub-task solving. For instance, the Option-critic Framework introduces an architecture capable of learning higher and lower-level policies without

needing prior knowledge of sub-goals Bacon et al. [2017]. HRL has demonstrated superior performance in various domains, including long-horizon games, continuous control problems Claure et al. [2020] and fairness in human-in-the-loop IoT Elmalaki [2021]. In FAIRO, we will decompose the fairness problem into sub-tasks over smaller time horizons and exploit the options framework to solve these sub-tasks.

2.3.2.4 Contributions of FAIRO

The contributions of FAIRO can be summarized as follows:

- **Fairness from the lens of equity:** We tackle the fairness problem in sequential decision-making systems within HITL environments by addressing the notion of equity.
- **FAIRO:** We propose FAIRO, a novel algorithm designed for fairness-aware sequential-decision making in HITL adaptation. Our approach leverages the Options RL framework to effectively incorporate fairness.
- **Generalization to different HITL application setups:** We extend FAIRO to cater to three types of HITL application setups. These setups involve multiple humans sharing the application space and being impacted by: (1) global numerical adaptation decisions, (2) shared global resources, and (3) shared global categorical adaptation decisions.
- **Evaluation on multiple HITL applications:** We conduct comprehensive evaluations of FAIRO on three different HITL applications to demonstrate its generalizability and compare with previous work in the literature.

The remainder of this section is structured as follows: Section 2.3.3 summarizes the Options framework, which serves as the foundation for our proposed approach. Section 2.3.4 details how to incorporate fairness considerations into the decision-making process. The subsequent sections focus on evaluating our proposed approach, FAIRO, in three distinct application domains.

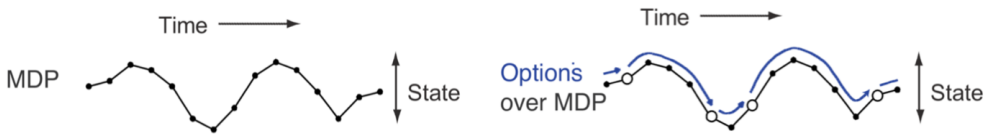


Figure 2.7: The state trajectory of an MDP with small discrete-time transitions. Options enable overlaid larger abstracted discrete events.

2.3.3 Options framework for temporal abstraction

Markov Decision Process (MDP) is widely employed for modeling sequential decision making. Various methods are utilized to solve MDPs and obtain the optimal Markov decision chain, including dynamic programming and reinforcement learning (RL). RL is particularly used when the transition probabilities within the MDP are unknown. Within the discrete-time finite MDP setting, the standard RL framework can be applied. In particular, an *agent* engages with an *environment* that is modeled as an MDP at discrete time steps, denoted as $t = 0, 1, 2, ...$ At each time step t, the agent observes the current state of the environment, denoted as $s_t \in \mathcal{S}$, and selects an action $a_t \in \mathcal{A}$ based on this observation. This action leads to a transition to the next state, s_{t+1}, and yields a reward value, $r_t \leftarrow \mathbb{R}$, associated with this transition. By engaging in this interaction, the agent learns a policy $\pi(s, a)$ that guides its decision-making, aiming to select the best action a for each state s to maximize the expected total reward over sequential decision actions.

The options framework was first introduced by Sutton et al. Sutton et al. [1999] to generalize primitive actions to include temporally extended courses of lower-level action. In particular, the term options represents a temporal abstraction of the lower-level actions in the MDP. A pictorial figure of options over MDP is shown in Figure 2.7. An MDP's state trajectory comprises small, discrete-time transitions, whereas the options enable an MDP to be abstracted and analyzed in larger temporal transitions.

Option o within the option set $\mathcal{O}$ consists of three main components: a policy $\pi(a|s, o)$ for selecting actions within option o, an initiation set $\mathcal{I} \subseteq \mathcal{S}$, a termination condition β. An

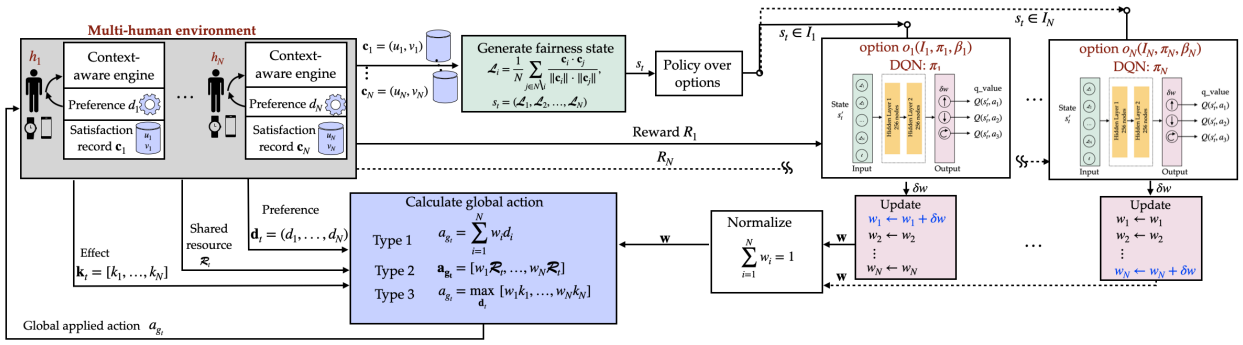


Figure 2.8: FAIRO for fairness-aware framework in Human-in-the-Loop systems using options framework. FAIRO is designed for three types of applications: **Type I:** one global action based on numerical demands affecting multiple humans, **Type II:** one shared resource distributed over multiple humans, and **Type III:** one global action based on categorical preferences.

option o:$(\mathcal{I}, \pi, \beta)$ is available to be selected by the agent in state s_t if and only if $s_t \in \mathcal{I}$. If the option is selected, actions are selected according to the option policy π until the option terminates according to the termination condition β. When the option terminates, the agent can select another option. This definition of options makes them act as much like actions while adding the possibility that they are temporally extended[5].

In FAIRO, the rationale behind employing the options framework to achieve fairness in a multihuman setting stems from the inherent limitations imposed by an option's initiation set $\mathcal{I}$ and termination condition β. These constraints confine the applicability of an option's policy, π, to a subset defined by $\mathcal{I}$ rather than encompassing the entire state space $\mathcal{S}$. Consequently, options can be viewed as a means of achieving **fairness subgoals**, wherein each option's policy is adapted to enhance the attainment of its specific **subgoal**, thereby contributing to the overall fairness of the decision-making agent. The dynamic nature of the multihuman environment necessitates diverse fairness policies at different temporal instances.

[5]Options framework can be extended to include policies over options. When multiple options are available to the agent at s_t, the agent can learn which option to select using the policy over options. We consider the policy over options to be a fixed policy, and the initiation sets of all options are disjoint sets.

2.3.4 FAIRO: Fairness using Options framework

We exploit options framework to design FAIRO to achieve fairness in sequential decision-making agents in multihuman environment Sutton et al. [1999]. As seen in Figure 2.8, the agent interacts in sequential discrete-time steps with an environment that has N humans $(h_1, h_2, \ldots, h_N)$ through observing their preferences or their desired adaptation actions $(d_1, d_2, \ldots, d_N)$ and the current fairness state of the environment s_t. Guided by the current fairness state s_t, the agent selects an appropriate option o_t from the set of N available options $\mathcal{O}$. The chosen option o_t then determines a lower-level action based on its specific option policy π_o, resulting in a global action a_{g_t} that is applied to the shared environment. This global action subsequently modifies the current fairness state, and the agent receives a reward r_{t+1}. This reward is utilized to refine the option policy. In the following subsections, we provide a detailed description of each module within the FAIRO framework.

2.3.4.1 Fairness state space $\mathcal{S}$

Our approach to viewing fairness from the lens of equity is by using a fairness state that encompasses the history of the positive and negative effects of the global decision action.

Satisfaction history records $\mathbf{c}_i$ Fairness state s_t is inferred from the history of the *satisfaction* of each human. To model the satisfaction of the human h_i, we keep a history record for each human:

$$\mathbf{c}_i = (u_i, v_i), \text{where } i \in \{1, 2, \ldots, N\}. \tag{2.14}$$

The value $u_i \in \mathbb{R}$ represents a record of the number of times the human h_i was unsatisfied by the applied global action a_g. In contrast, $v_i \in \mathbb{R}$ represents a record for the number of times the human h_i was satisfied by the applied global action a_g.

At time step t, every human h_i has a desired adaptation action d_{i_t}. For example, a human

may prefer a particular temperature setpoint to HVAC system (Heating, ventilation, and air conditioning) in their room for thermal comfort that matches their physical activities, such as sleeping, domestic work, or sitting. Based on the difference in the values of d_{i_t} and a_{g_t}, the record $\mathbf{c}_i$ is updated to capture whether the human was satisfied or unsatisfied. For example, if this difference is within a threshold τ then we consider the human h_i is satisfied and increment v_i by a value δ.

$$\mathbf{c}_i = \begin{cases} (u_i, v_i + \delta) & \|d_{i_t} - a_{g_t}\| \leq \tau. \\ (u_i + \delta, v_i) & \|d_{i_t} - a_{g_t}\| > \tau. \end{cases} \tag{2.15}$$

After all the records $\mathbf{C} = (\mathbf{c}_i, i = 1, 2, ..., N)$ are updated, they are normalized to a unit vector. Choosing the value τ is application dependent; however, the value δ needs to be less than 1 and small enough to ensure that the unit vector direction $\mathbf{C}$ does not change drastically. Hence, we choose δ to be 0.01.

Ideally, these records $\mathbf{c}_i$ should be $(0, 1)$ indicating that the global adaptation action a_g meets the preferences of the human over time. However, as we mentioned earlier, these preferences may conflict with humans sharing the same environment. Hence, the same a_{g_t} may be perceived by one human as meeting their preference (increasing v) and by another human as not meeting theirs (increasing u).

A pictorial visualization of $\mathbf{C}$ is shown in Figure 2.9. Each $\mathbf{c}_i$ can be represented as a $2D$ vector within a unit circle. We show two examples for $\mathbf{c}_i$ for three humans where h_3 has $\mathbf{c}_3$ closer to the v axis compared to the other two humans (Figure 2.9-left) versus the case when h_3 has $\mathbf{c}_3$ closer to the u compared to the other two humans (Figure 2.9-right). Figure 2.9 shows an example of a relatively unfair situation, where h_3 is either treated most of the time favorably (Figure 2.9-left) or unfavorably (Figure 2.9-right).

It is worth mentioning here that $\mathbf{C}$ captures the history of the effect of the trajectory of

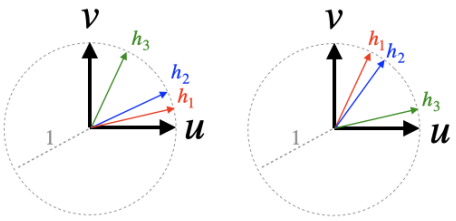


Figure 2.9: Satisfaction history record $\mathbf{C}$ is represented as $2D$ unit vector, where the two components u and v represent history of "unsatisfied" and "satisfied" respectively.

sequential adaptation action on the shared environment. Hence, the intuition is to tune the global action in the next time step $a_{g_{t+1}}$ to either decrease the $focus$ on considering the preferences of h_3 (Figure 2.9-left) or vice versa (Figure 2.9-right). However, as mentioned in Section 2.3.1, the same action affects all the humans sharing the same environment.

Fairness state s_t We use the geometric intuition in Figure 2.9 to design our fairness state s_t to compare the directions of all N records in $\mathbf{C}$. Ideally, we would like to have all $\mathbf{c}_i$ as close as possible to each other. Hence, we define s_t to capture how close each $\mathbf{c}_i$ is to the other $N \backslash i$ records. Hence, we define (s_t) as follows:

$$\mathcal{L}_i = \frac{1}{N} \sum_{j \in N \backslash i} \frac{\mathbf{c}_i \cdot \mathbf{c}_j}{\|\mathbf{c}_i\| \cdot \|\mathbf{c}_j\|}, \qquad \mathcal{L}_i \in [0, 1] \subset \mathbb{R} \tag{2.16}$$

$$s_t = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N), \qquad s_t \in \mathcal{S} = [0, 1]^N \tag{2.17}$$

In particular, $\mathcal{L}_i$ represents the closeness of record $\mathbf{c}_i$ to the rest of the records using the average of the cosine of the angle between pair of vectors. Hence, if the cosine value between two vectors is 1, they coincide. Since the values of $\mathbf{c}_i$ can only be positive and are normalized to a unit vector, the minimum cosine value between these vectors is 0, indicating that they are far from each other (at 90°)[6]. Equation 2.17 represents s_t which holds all the values of $\mathcal{L}_i$. Ideally, from our fairness point of view, the goal state s_t should be $(1, 1, \ldots 1)$, which indicates that all $\mathbf{C}_h$ have the same direction, meaning that the history of the satisfaction and unsatisfaction for all the humans are close.

[6] $\mathbf{c}_i \in \mathbb{R}, \mathcal{L}_i$ is unlikely to reach 0 but can decrease to a very small value ϵ.

2.3.4.2 Initiation set $\mathcal{I}$ and fairness subgoals

While the ultimate goal is to learn a policy that can achieve the goal state $s_t = (1, 1, \dots 1)$, this is challenging since it is a huge state space. Accordingly, the intuition behind exploiting the options framework is to divide this goal into smaller subgoals where we learn over a subset of states or the initiation set ($\mathcal{I} \subseteq \mathcal{S}$) as explained in Section 2.3.3. We divide $\mathcal{S}$ into N initiation sets $\mathcal{I}_i$ where $i \in \{1, 2, \dots, N\}$, such that $\mathcal{I}_i$ contains all the states with $\mathcal{L}_i$ as the minimum value.

$$\mathcal{I}_i = \{s_t = \{\mathcal{L}_1, \dots, \mathcal{L}_N\} \in \mathcal{S} | \mathcal{L}_i = \min(s_t)\} \tag{2.18}$$

Specifically, this means that each initiation set $\mathcal{I}_i$ considers only the states where h_i has received unfair adaptation either favorably or unfavorably. For example, both cases in Figure 2.9 are considered unfair state where $\mathcal{L}_3$ is less than $\mathcal{L}_1$ and $\mathcal{L}_2$.

2.3.4.3 Termination State β

Each option o_i terminates when the current state s_t reaches a terminal state for this option. Hence, in FAIRO, the set of terminal states for o_i is when $\mathcal{L}_i$ is no longer the minimum value in s_t.

$$\beta_i = \{s_t = \mathcal{L}_1, \dots, \mathcal{L}_N\} \in \mathcal{S} | \mathcal{L}_i \neq \min(s_t)\} \tag{2.19}$$

Intuitively, this means that each option o_i will run to improve the value of $\mathcal{L}_i$ until it is no longer the minimum value which is the fairness subgoal for this option. This will trigger a new initiation set I and this option terminates and a new option starts to achieve another subgoal: improving $\mathcal{L}_i$.

2.3.4.4 Global action of different HITL applications

As shown in Figure 2.8, every human (h_i) has a desired preference (d_i). However, only one action a_g is chosen to be applied to the shared environment. In FAIRO, we identify three

types of applications:

- **(Type I) Shared numerical global action:** The desired preferences d_i have numerical values and the global action a_g is a numerical value. We design each option to take a weighted sum of these N preferences. These weights represent the contribution of each human preference d_i in the applied global action a_{g_t}. Hence, $a_{g_t} = \sum_{i=1}^{N} w_i d_i$, where $w_i \in [0, 1]$. We will show an instant of this type in a simulated smart home application in Section 2.3.5.
- **(Type II) Shared global resource:** The desired preferences d_i have numerical values. There is one shared time-varying resource $\mathcal{Z}$ and it has to be distributed. The global action $\mathbf{a}_{\mathbf{g}_t}$ is the weighted share of this resource $\mathcal{R}_t$ dictated by their desired preferences. Hence, $\mathbf{a}_{\mathbf{g}_t} = [w_1\mathcal{Z}_t, w_2\mathcal{Z}_t, \ldots, w_N\mathcal{Z}_t]$, where $w_i \in [0, 1]$. We will show an instant of this type in a simulated water distribution application in Section 2.3.6.
- **(Type III) Shared categorical global action:** The desired preferences d_i have categorical values, with the global action a_g selecting one of these categorical values, which represents the maximum weighted effect of applying d_i on all other $N - i$ humans, where the effect k_i can be estimated from the context-aware engine, resulting in

$$\mathbf{a}_{\mathbf{g}_t} = \arg\max_{d_i}[w_1k_1, w_2k_2, \ldots, w_Nk_N], where w_i \in [0, 1]. \tag{2.20}$$

  We will show an instant of this type in a smart education application in Section 2.3.7.

2.3.4.5 Learning the option policy (π_i)

Each option policy π_o learns the appropriate $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ for the global action a_{g_t} for every state $s_t \in \mathcal{I}_i$ to reach the termination condition β_i such that $\sum_{i=1}^{N} w_i = 1$. These weights are continuous values and learning them for every $s_t \in \mathcal{I}_i$ is challenging. Hence, we opt for a simpler design of using Deep Q-Network (DQN) to reduce the search space for the appropriate weights as detailed below.

Deep Q-Network (DQN) DQN is a reinforcement learning algorithm that utilizes deep neural networks in combination with Q-learning within Markov Decision Processes (MDP), estimates the action-value function $Q(s, a)$, representing expected rewards for actions in states. This function is typically represented by a neural network with state inputs and estimated action values as outputs, and at each time step t, the DQN agent uses ε-greedy policy to select action, updating the Q-function based on observed states, rewards, and the next state using an update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)), \tag{2.21}$$

where α is the learning rate, γ is the discount factor, and $\max_a Q(s_{t+1}, a)$ is the estimated maximum action-value in the next state s_{t+1}. DQN aims to learn a policy that maximizes the cumulative reward over time in a given environment. It combines Q-learning with deep learning, allowing the agent to handle high-dimensional observations and non-linear function approximations. Figure 2.10 shows a pictorial illustration for the design of the DQN in every option o_i in FAIRO.

Input to Option DQN (s'_t) Every option o_i runs a DQN where the input is the s_t. However, to differentiate between the two cases shown in Figure 2.9 where both $\mathcal{L}_3$ is the minimum value in s_t, we append to s_t the relative location of $\mathbf{c}_3$ with respect to the $\mathbf{c}_1$ and $\mathbf{c}_2$.

$$s'_t = (s_t, l), \text{ where } s_t \in \mathcal{I}_i, \quad l = \begin{cases} 1, & v_i = \max_v(s_t). \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

This means that if $\mathcal{L}_i$ is the minimum in s_t, option o_i will run since $s_t \in \mathcal{I}_i$. The input to the DQN in option o_i will indicate whether $\mathcal{L}_i$ has the minimum value (unfair situation) because h_i received favorable treatment relative to all $h_{N\setminus i}$ (*i.e.,* $\mathbf{c}_i$ has a high v_i component) and in

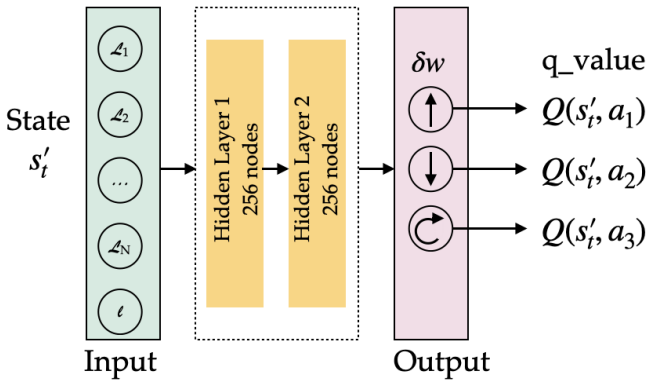


Figure 2.10: DQN for option o_i. The input is the state s'_t. The output is the `q_value` for the possible three actions of weight adjustment δw.

this case $l = 1$ in Equation 2.22, or $l = 0$ otherwise.

Output from Option DQN (δw_i) To reduce the action space of the DQN since we need to learn N weights with continuous values $[0, 1]$, we designed the output from the DQN in option o_i to focus only on adjusting w_i instead of all the N weights $\mathbf{w}$. In particular, as shown in Figure 2.10, the output from the DQN are the Q-values ($Q(s, a)$) which decides to select between three actions to (a_1) increase the weight $w_i \uparrow$, (a_2) decrease the weight $w_i \downarrow$, or (a_3) keep the same weight $w_i \circlearrowleft$. Hence, the output from DQN (which is the selected action of the DQN as explained in Section 2.3.4.5) is the weight adjustment for w_i by $\delta w_i = (+\delta, -\delta, 0)$ where the value δ determines how fast the DQN for option o_i changes the weight w_i.

$$w_i \leftarrow w_i + \delta w_i, \qquad \sum_{i=1}^{N} w_i = 1. \tag{2.23}$$

Since all the weights $\mathbf{w}$ have to be normalized to 1, all the weights will be adjusted accordingly. Using this design, we choose between 3 possible adjustments for weight w_i instead of the whole space $[0, 1]$ and instead of all the N weights.

Option DQN reward ($\mathcal{R}_i$) Each option DQN learns the appropriate policy $\pi_i(s'_t, \delta w_i)$, which is the right weight adjustment δw_i at state s^i_t. The DQN learns this policy through a notion of a feedback reward as explained in Section 2.3.4.5. As each option o_i aims at

increasing the fairness subgoal of enhancing $\mathcal{L}_i$ while improving the performance of the application, the reward function $\mathcal{R}_i$ can be expressed with two terms; a fairness term $\mathcal{F}$, and a performance term $\mathcal{P}$ using a trade-off parameter $\zeta \in [0, 1]$.

$$\begin{aligned}
\mathcal{R}_i &= \zeta\mathcal{F}_i + (1-\zeta)\mathcal{P}_i, \in [-1, 1] \subset \mathbb{R} \\
\mathcal{F}_i &= \text{absolute fairness} + \text{option improvement} \\
&= (2 * \mathcal{L}_{i_t} - 1) + f(\mathcal{L}_{i_{t-1}}, \mathcal{L}_{i_t}), \in [-1, 1] \subset \mathbb{R} \\
\mathcal{P}_i &= \text{application dependent}, \in [-1, 1] \subset \mathbb{R}
\end{aligned} \tag{2.24}$$

The fairness $\mathcal{F}_i$ considers the current value of $\mathcal{L}_i$, which we call the "absolute fairness", and the "improvement in the fairness" value of $\mathcal{L}_i$ from last time step for this particular option. It is a function (f) of the current value of $\mathcal{L}_{i_t}$ and the value from last time step $\mathcal{L}_{i_{t-1}}$ as shown in Equation 2.24. The overall FAIRO algorithm is listed in Algorithm 1.

2.3.5 Application Type I: Smart Home HVAC

Recent literature focuses on enhancing human satisfaction in smart heating, ventilation, and air conditioning (HVAC) systems by employing reinforcement learning (RL) techniques to adjust the set-point based on human activity and preferences Elmalaki [2021]. These HITL systems consider the current state and individual preferences, such as body temperature changes during sleep or physical activity. To evaluate FAIRO in **Type I** applications, we consider a setup where multiple humans share a house with a single HVAC system, and their activities determine individual desired setpoint.

2.3.5.1 House-Human Physical Model

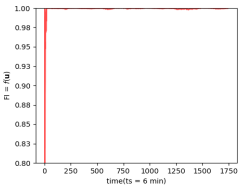
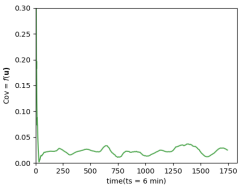
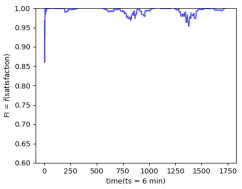
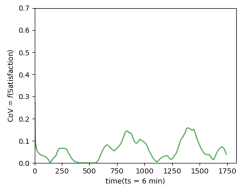
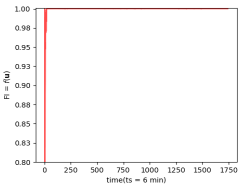
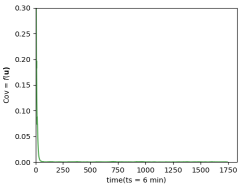
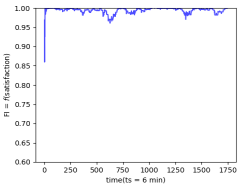
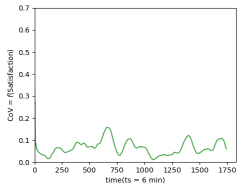
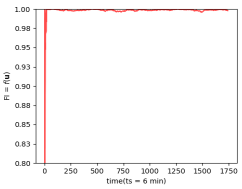
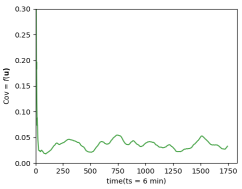
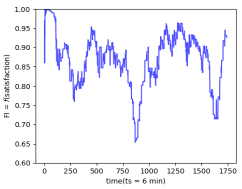
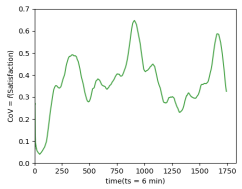
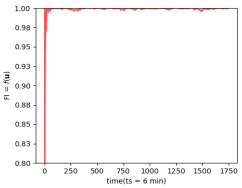
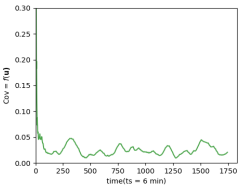
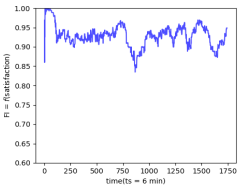
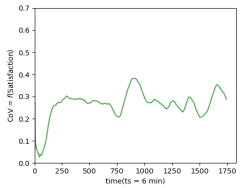
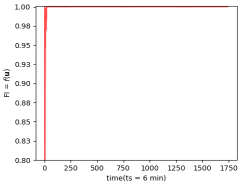
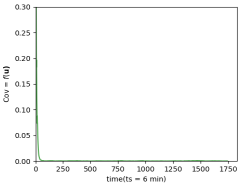
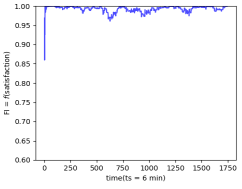
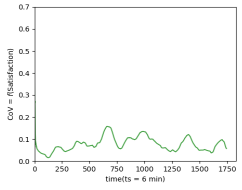
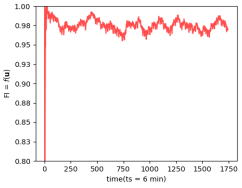
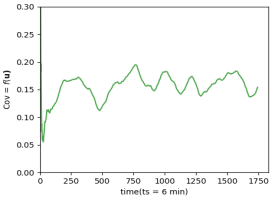
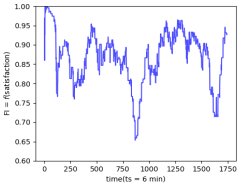
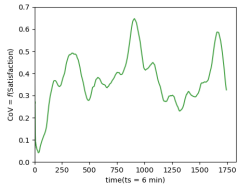
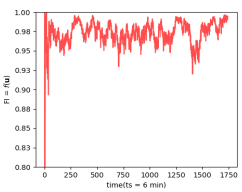
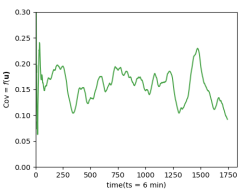
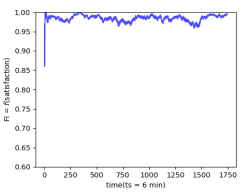
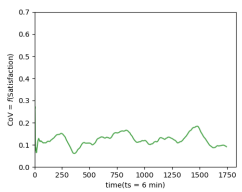
We used a thermodynamic model of a house incorporating the house's shape and insulation type. To regulate indoor temperature, a heater and a cooler with specific flow temperatures ($50°c$ and $10°c$) were employed. A thermostat maintained the indoor temperature within $2.5°c$ around the desired set point. An external controller controls the setpoint. The hu-

man was modeled as a heat source, with heat flow dependent on the average exhale breath temperature (EBT) and the respiratory minute volume (RMV). These parameters depend on human activity Carroll [2007]. We simulated three humans with four activities: sleeping, relaxing, medium domestic work, and working from home. The different activity schedules depicted in Figure A.1-Left in the Appendix. The humans were simulated in separate rooms, each exhibiting unique behavioral patterns: (1) h_1 followed an organized and repetitive weekly routine, (2) h_3 had a more random and unpredictable life pattern, and (3) h_2 displayed intermediate randomness, alternating between sleeping, being away from home, domestic activities, and relaxation. The Mathworks thermal house model was extended to include a cooling system, a human model, and an external controller running FAIRO[7].

2.3.5.2 Context-aware engine

A context-aware engine estimates the desired action d_i per human h_i in a smart home. The desired action d_i can be obtained through fixed policy configuration or learned policy. To focus on our main contribution and not on designing a new context-aware engine, we leverage existing RL-based approaches for estimating the desired HVAC setpoint d_i based on activity and thermal comfort Taherisadr et al. [2023]. The desired setpoints for the considered activities are domestic activity (72°F), relaxed activity (77°F), sleeping (62°F), and work from home (67°F). These setpoints aim to enhance thermal comfort. Thermal comfort is assessed using Prediction Mean Vote (PMV) on a scale from very cold (-3) to very hot ($+3$). Optimal indoor thermal comfort falls within the recommended range of $[-0.5, 0.5]$, as per the ISO standard ASHRAE 55 ASHRAE/ANSI Standard 55-2010 American Society of Heating, Refrigerating, and Air-Conditioning Engineers [2010].

Table 2.3: Comparison between all the different five approaches of FinA, Mean approach, and Round Robin

	$FI_{\mathbf{u}}$	$CoV_{\mathbf{u}}$	$FI_{\mathbf{SR}}$	$CoV_{\mathbf{SR}}$
Approach I				
Approach II				
Approach III				
Approach IV				
Approach V				
Mean				
Round Robin				

Algorithm 1 FAIRO algorithm

[1] Humans $\mathcal{H} = (h_1, \ldots, h_N)$ Satisfaction records $\mathbf{C} = (\mathbf{c_1}, \ldots, \mathbf{c_N}), \mathbf{c}_i = (u, v) \in \mathbb{R}^2$ States $s \in \mathcal{S} =]0, 1]^N \subset \mathbb{R}^N$ Initiation sets $\mathcal{I}_i \in \mathcal{I} = \{s \in \mathcal{S}\}$ Options $\mathcal{O} = (o_1, \ldots, o_N), o_i = (\mathcal{I}_i, \pi_i, \beta_i)$ Application type $T = \{Type1, Type2, Type3\}$ Run-FAIRO True $\mathbf{d}_t \leftarrow$ context-aware-engine $(\mathcal{H})$ $T == Type2$ $\mathcal{R}_t \leftarrow$ get-current-available-resource() $T == Type3$ $\mathbf{k}_t \leftarrow$ get-current-effects$(\mathbf{d}_t)$ $s_t \leftarrow$ get-fairness-state$(\mathbf{C})$ $s'_t \leftarrow$ append-state(s_t) $o_t \leftarrow$ choose-option (s_t) $o_t \in \mathcal{O}$ $\delta w \leftarrow$ run-option(o_t) option policy $\pi_o(s_t, \delta w)$ $\mathbf{w}_t \leftarrow$ update-normalize-weights(δw) $a_{g_t} \leftarrow$ calculate-global-action$(\mathbf{d}_t, \mathbf{w}_t, \mathcal{R}_t, \mathbf{k}_t)$ Apply global action a_{g_t} on the shared environment $\mathcal{R}_t \leftarrow$ receive-reward () $\pi_o(s_t, \delta w) \leftarrow$ Update-option-policy (R_t) $\mathbf{C} \leftarrow$ Update satisfaction records $(\mathbf{d}_t, a_{g_t})$

Figure 2.11: Fairness state and satisfaction across all approaches in application 1 showing the probabilities of equal opportunity and equalized odds.

2.3.5.3 Evaluation

We compare 5 different approaches including one of the state-of-the-art approaches FaiRIoT Elmalaki [2021]:

- **FAIRO:** The global applied action is $a_{g_t} = \sum_{i=1}^{3} w_i d_i$. The reward per option i is as explained in Equation 2.24. We set $\zeta = 0.5$. As for the performance term $\mathcal{P}_i$ in the reward, we assign a high reward when the PMV falls in the acceptable range $[-0.5, 0.5]$.

[7] While more complex simulators like EnergyPlus exist, we opted for a simpler model to assess FAIRO.

Further, we used the values of the satisfaction counters $\mathbf{c}_i = (u_i, v_i)$ as an indication of the performance $\mathcal{P}_i$ since their values are correlated to the desired temperature d_i which maps to the best PMV. The value $\mathcal{P}_i$ is then normalized to be $[-1, 1]$:

$$
\begin{aligned}
& f(\mathcal{L}_{i_{t-1}}, \mathcal{L}_{i_t}) = sign(\mathcal{L}_{i_{t-1}} - \mathcal{L}_{i_t}) \times Z \\
& Z = \begin{cases} 0 & |\mathcal{L}_{i_{t-1}} - \mathcal{L}_{i_t}| \in]0, 0.001] \\ 0.25 & |\mathcal{L}_{i_{t-1}} - \mathcal{L}_{i_t}| \in]0.001, 0.005] \\ 0.5 & |\mathcal{L}_{i_{t-1}} - \mathcal{L}_{i_t}| \in]0.005, 0.01] \\ 0.75 & |L_{i_{t-1}} - \mathcal{L}_{i_t}| \in]0.01, 0.015] \\ 1 & |L_{i_{t-1}} - \mathcal{L}_{i_t}| > 0.015 \end{cases} \\
& \mathcal{P}_i = 0.2 \frac{v_i}{u_i + v_i} + 0.8 f(\text{PMV}), \in [-1, 1] \subset \mathbb{R}
\end{aligned}
\tag{2.25}
$$

- **Average approach:** The setpoint is the mean value of desired setpoints of all rooms. Hence, $a_{g_t} = \frac{1}{3} \sum_{i=1}^{3} d_i$.
- **Equality using round robin (RR):** The setpoint is selected from one of the desired setpoints of all rooms in a rotation. The intuition of this approach is to compare with the case where we give every room the same opportunity to use its desired setpoint across time (equality). Hence, for 3 humans, $a_{g_1} = d_1, a_{g_2} = d_2, a_{g_3} = d_3, a_{g_4} = d_1, \ldots$, etc.
- **No subgoals using 1 DQN:** The setpoint is calculated using a single model of 1 DQN structured as Figure 2.10. The inputs are the three values of the fairness state ($s_t = (\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3)$). The intuition of this approach is to evaluate whether we can achieve the overall fairness goal without the need for subgoals that FAIRO provides. The reward for the single DQN is the average reward value across all rooms:

$$
\begin{aligned}
& \mathcal{R} = 0.5\mathcal{F} + 0.5\mathcal{P}, \quad \text{where } N = 3 \in [-1, 1] \subset \mathbb{R} \\
& \mathcal{F} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i^N + f(\frac{1}{N} \sum_i \mathcal{L}_{i_{t-1}}, \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{i_t}), \\
& \mathcal{P} = 0.2 \frac{1}{N} \sum_{i=1}^{N} \frac{v_i}{u_i + v_i} + 0.8 \frac{1}{N} \sum_{i=1}^{N} f(PMV)
\end{aligned}
\tag{2.26}
$$

- **FaiRIoT Elmalaki [2021]:** The closest to our approach is FaiRIoT which uses hierarchical RL.

We investigated multiple evaluation metrics to evaluate the fairness across these three rooms; 1) the values of the fairness state ($s_t = (\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3)$, 2) a satisfaction metric, 3) PMV, and 4) the covariance cv of the learnt weights.

Fairness state and group fairness definition As explained in Section 2.3.4.1, the best fairness state should be $s_t = (\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3) = (1, 1, 1)$. To measure s_t, we need to use the satisfaction history counters by updating them per to Equation 2.15. In this application, we set the threshold τ to be 2.5. Figure 2.11 shows the fairness states results with 15k samples equivalent to 62.5 simulated days.

To better get insights on how the different methods compare with respect to the values of s_t, in every time sample, we report the maximum value of $\mathcal{L}$ in s_t (Figure 2.11-first row) and the minimum value of $\mathcal{L}$ in s_t (Figure 2.11-second row). We plot the values $\mathcal{L}_1$ for Room 1(red), $\mathcal{L}_2$ for Room 2(blue), and $\mathcal{L}_3$ for Room 3(green). The results for using only 1 DQN are unstable, and have fluctuations in the fairness state values. This result is expected for 1 DQN since the fairness goal was not divided into sub-goals, unlike what the options framework provided.

The commonly used metrics for group fairness are *equal opportunity* and *equalized odds*. Although these metrics are typically applied to binary classification tasks, we adopt their original definitions to compare FAIRO with the other approaches.

• ***Equal opportunity*** aims to ensure that individuals from different groups have an equal chance or probability of experiencing positive outcomes or receiving beneficial treatment or resources. To assess the performance of FAIRO, we analyze the results presented in Figure 2.11 by examining the reported $\max_{\mathcal{L}} s_t$ values. Specifically, we compare the proba-

bilities of different rooms M_i having the highest $\max_{\mathcal{L}} s_t$ values. For *equal opportunity*, these probabilities need to be close, as denoted in Equation 2.27.

$$p(M_i == \texttt{find}(\max_{\mathcal{L}} \ s_t)) \approx p(M_{N\setminus i} == \texttt{find}(\max_{\mathcal{L}} \ s_t)) \tag{2.27}$$

As observed in Figure 2.11-first row, these probabilities are 34%, 29.4%, and 36.6% for Room M_1, M_2, and M_3 respectively, which are closer in values compared with the other approaches. In particular, using FAIRO, the average absolute difference between the probabilities of *equal opportunity* across the 3 rooms is reduced by 57.0%, 54.8%, and 35.8% from Average Approach, RR, and 1 DQN respectively. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by 49.2% on average.**

• ***Equalized odds*** focuses on the balance between positive and negative outcomes across different groups. Hence, for the negative outcomes, we can examine the probabilities of different rooms having the $\min_{\mathcal{L}} s_t$ values and assess whether they are approximately equal.

$$p(M_i == \texttt{find}(\min_{\mathcal{L}} \ s_t)) \approx p(M_{N\setminus i} == \texttt{find}(\min_{\mathcal{L}} \ s_t)) \tag{2.28}$$

As observed in Figure 2.11-second row, these probabilities are $34.5\%, 34.9\%$, and 30.7% for Room M_1, M_2, and M_3 respectively, which are closer in values compared with the other approaches. In particular, using FAIRO the average absolute difference between the probabilities of *equalized odds* across the 3 rooms is reduced by 47.8%, 35.8%, and 41.3% from Average Approach, RR, and 1 DQN respectively. **Across all the approaches, FAIRO improves the *equalized odds* fairness by 41.63% on average.**

Satisfaction performance Figure 2.11-third row shows the satisfaction values ($\frac{v}{v+u}$) for all methods. FAIRO achieves close satisfaction values across the three rooms over time. Average and RR approaches have stable but different satisfaction values across rooms while 1 DQN shows more fluctuations per room. We focused on samples after FAIRO converges

($12k$ to $15k$) and examined the satisfaction value histograms. We report the Jensen-Shannon Divergence (JSD) of these histograms[8]. FAIRO has the lowest average JSD (0.013), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO reduces JSD by 92.06% on average.** More details are shown in Figure A.1-right in Appendix.

To gain further insights into human satisfaction, we plot the covariance of temperature differences between the applied setpoint a_g and the desired temperature d_i for the three humans, and Figure 2.12 shows that FAIRO exhibits the lowest cv of 0.04, indicating better fairness.

PMV results FAIRO achieves the second lowest average JSD and a comparable variance on PMV results, which indicates FAIRO can improve fairness without hurting the PMV performance. RR achieves the lowest PMV JSD average since every time step one of the rooms can get exactly its desired temperature. Hence, the 3 rooms can get almost identical PMV performance in the long term. **Across all the approaches, FAIRO's PMV JSD is reduced by 13.7% on average.** More details are shown in Figure A.1.

Comparison with the State-of-the-Art FaiRIoT Elmalaki [2021] FaiRIoT uses a notion of utility $u_{h_t} = \frac{1}{t}\sum_{j=0}^{t}\frac{j}{t}w_{h_j}$ which is the average weight assigned by a layer called "Mediator RL" for a particular human h over a time horizon $[0:t]$, where the factor $\frac{j}{t}$ is used to give more value to the recent weights learnt by the Mediator RL over the ones in the past. FaiRIoT measures the fairness of the Mediator RL using the coefficient of variation (cv) of the human utilities: $cv = \sqrt{\frac{1}{n-1}\sum_{h=1}^{n}\frac{(u_h-\bar{u})^2}{\bar{u}^2}}$, where $\bar{u}$ is the average utility of all humans. The Mediator RL is said to be more fair if and only if the cv is smaller. Accordingly, we compare the cv in FaiRIoT and FAIRO in Figure 2.13. FAIRO achieves cv around 0.15, while FaiRIoT cv is larger than 0.6. **Hence, FAIRO improves the fairness where *cv* is**

[8]The Jensen–Shannon divergence (JSD) is a symmetric measure of similarity between two probability distributions, always non-negative, with 0 denoting identical distributions and any value above 0 indicating differences.

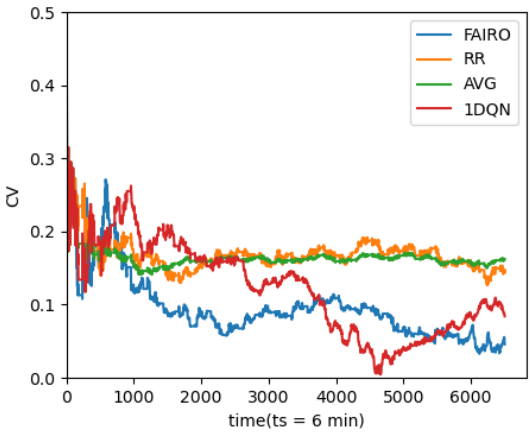


Figure 2.12: Coeff. of variation (cv) of the temperature differences.

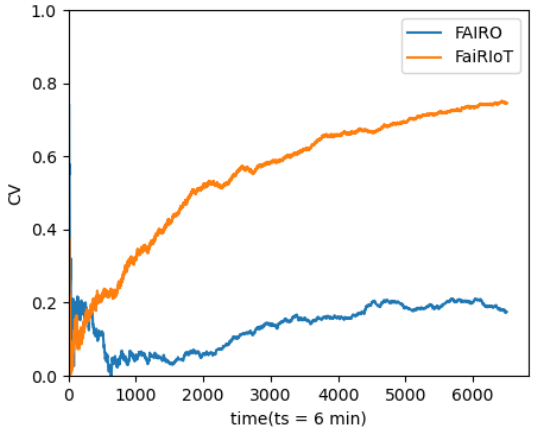


Figure 2.13: Coeff. of variation (cv) of the utility u_t over $\mathbf{w}$.

reduced by 75%.

2.3.6 Application Type II: Water supply application

In the context of global climate change and rapid population growth, Water Demand Models (WDMs) are crucial for gaining insights into water consumption behavior and forecasting demand, aiding decision-making in water distribution system (WDS) operation policies and infrastructure planning. However, managing limited water resources while ensuring equitable distribution among households remains a challenge, with various pricing and non-pricing policies explored in the literature Sechi et al. [2013]. For evaluating FAIRO in application Type II, it is assumed that WDM estimates desired water demand for three households from a time-varying, insufficient shared water resource.

2.3.6.1 Household-Water Demand Physical Model

We utilized a WDM to model water demand based on residents' activity patterns, matching the activity patterns presented in Section 2.3.5.1 to water demand behavior Philadelphia Government [2023]. The water demand patterns for three households are illustrated in Figure A.4-Left in the Appendix. Each household has a water reservoir $\mathcal{T}$ and aims to meet its specific water demand d_i, with the shared water resource ($\mathcal{Z}_t$) being variable and

insufficient to satisfy all demands. Assuming the availability of the WDM, we set $\mathcal{Z}_t$ to 1.5 times the maximum demand among the three profiles, with the water demand profiles being independent due to distinct human activities. We expanded the Mathworks water supply physical model to accommodate these household water demand profiles and multiple houses MATLAB [2023].

2.3.6.2 Context-aware engine

Our main contribution is not designing a new water consumption behavior or forecasting demand models, hence, the context-aware engine provides the desired action for every household h_i, which is the current water demand d_i based on the water demand pattern. This demand is supposed to be satisfied via the water supply s_i from the shared limited resource $\mathcal{Z}_t$ and the current reserved water in the household water reservoir $\mathcal{T}_i$. Hence, we define the application performance as the percentage of balancing the demand and the supply.

$$\text{Balance Rate (BR)} = \frac{\text{supply } s_i + \text{reserve } \mathcal{T}_i}{\text{demand } d_i} \tag{2.29}$$

2.3.6.3 Evaluation

We compare the same methods as in the first application.

- **FAIRO:** The supply s_i for each household h_i is calculated by FAIRO. In particular, s_i receives $w_i \mathcal{Z}_t$ as explained in Section 2.3.4.4 where the global action a_{g_t} is the weighted distribution of this resource $\mathcal{Z}_t$. The reward is the same as explained in the first application (Equation 2.25), where performance $\mathcal{P}_i$ is based on the BR ($\mathcal{P}_i = 0.2\frac{v_i}{u_i+v_i} + 0.8f(\text{BR})$).
- **Average approach:** Each household h_i has the same amount of supply. Hence, $s_i = \frac{1}{3}\ \mathcal{Z}_t$. We further augment the average approach in this application by using a **Weighted Average approach**. In particular, each household supply is proportional to its demand. Hence, $s_i = \frac{d_i}{\sum_{i=1}^{3} d_i}\mathcal{Z}_t$.
- **Round Robin (RR):** Each time step, one of the households in a rotation will be guar-

anteed sufficient supply to cover its demand. Leftovers from the resource will be shared equally with all households. For example, at time $t = 1$, $s_1 = d_1$ while $s_2 = s_3 = \frac{Z_t - d_1}{2}$. We further augmented the RR to consider a **Weighted RR**. In this case, supplies are calculated similarly to RR, but the leftover water will be distributed proportionally to their demands as in the weighted average approach.

- **No subgoals using 1 DQN :** Supply for each household is calculated by 1 DQN structure as explained in Section 2.3.5.3.

We investigated multiple evaluation metrics as follows:

Fairness state and group fairness definition We simulated 62.5 simulated days equivalent to $15k$ samples. To measure s_t, we need to use the satisfaction history counters by updating them per Equation 2.15. In this application, we set the $\|d_i - (s_i + t_i)\| \leq \tau$ with τ equals 20% of the demand d_i which means BR is 80%. Similar to the analysis we did in application 1, FAIRO achieves *equal opportunity* probabilities 31.9%, 35.4%, and 32.7% for household #1, #2, and #3 respectively, which are closer in values than other methods. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by 26.38% on average.** As for *equalized odds* probabilities, FAIRO reports 34.4%, 30.1%, and 35.5% for households #1, #2, and #3 respectively, which are also closer in values compared to the other approaches. **Across all the approaches, FAIRO improves the *equalized odds* fairness by 32.1% on average.** More numerical details are shown in Figure A.2 in the Appendix.

Satisfaction performance We use $3k$ samples after FAIRO converge ($12k$ to $15k$) to examine the satisfaction value histograms. We report the Jensen-Shannon Divergence (JSD) of these histograms. FAIRO has the lowest average JSD (0.017), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO JSD is reduced by 91.06% on average.** More numerical details are shown in Figure A.4-Right in the Appendix.

Balance rate (BR) results We report the average number of samples that have $> 80\%$ BR across all households and average it over $3k$ samples. Samples have $> 80\%$ BR correspond to satisfactory samples. **Across all the approaches, FAIRO's average $> 80\%$ BR is improved by 46.66% on average.** More numerical details are shown in Figure A.4-Right in the Appendix.

2.3.7 Application Type III: Smart learning system

Monitoring human learning state and performance is crucial for assessing progress and personalizing instruction Terai et al. [2020]. Immersive technologies like virtual reality (VR) in education and workforce training show promise for improving learning experiences. However, over extended online or VR education periods, human performance can decline due to distractions, drowsiness, and fatigue Terai et al. [2020]. This application examines multiple users sharing the same educational environment, with a HITL learning system that adapts this environment by enabling an immersive VR experience to improve the learning experience.

2.3.7.1 Human-Learning VR Model

We used a real human dataset from recent literature studying VR's impact on learning Taherisadr et al. [2023]. The dataset describes human learning experience through three features: alertness, fatigue, and vertigo, resulting in 8 states with binary values (e.g., Alert: 1, Not Alert: 0). A HITL learning environment adapts based on these states using actions like (1) a_1 giving a break, (2) a_2 enabling VR, or (3) a_3 disabling VR.

We analyzed this dataset to categorize 15 participants into three groups based on VR tolerance: the most tolerant, those with some cybersickness, and those with the least VR tolerance. We model these groups' behavior using MDP models (see Figure A.3 in Appendix), with state 8 representing the best human state and state 1 the worst. Three humans, one

from each profile, were instantiated for a daily class schedule, with their states initialized to state 3 or state 1 randomly at the beginning of each day.

2.3.7.2 Context-aware engine

Unlike the previous two applications, this application has a non-numerical action space, as mentioned in Section 2.3.7.1. Hence, to use FAIRO we need to quantify this categorical action in terms of its effect as explained in Section 2.3.4.4. We exploit the MDP model in Figure A.3 in the Appendix to quantify these actions. In particular, as S_8 and S_0 encode the best and the worst state, respectively, we assign a value to each state linearly with S_8 as the highest value. Hence, the desired action d_i that can make a transition to a better state in the MDP has a high numerical value. However, every human may be in a different state. Hence, for every desired action d_i, we calculate the *effect* denoted as k_i of applying d_i on the *other* humans. Using the MDP models, if the desired action d_i from human h_i were to be applied in the shared environment, we check the state transition it will cause on all the other humans $h_{N\setminus i}$. The difference in the values of the two states in the transition is used to measure its *effect*. After applying an adaptation action, the human learning experience can be measured as the improvement in the human state values and the current human state value.

$$\text{Learn Exp. (LE)} = \text{state improvement} + \text{state value} \ \in [-1, 1] \tag{2.30}$$

2.3.7.3 Evaluation

- **FAIRO:** The global action $a_{g_t} = d_i$ is the desired action of the human$_i$ with the maximum weighted effect as explained in Section 2.3.4.4. The reward is the same as explained in the first application (Equation 2.25), where performance $\mathcal{P}_i$ is based on the learning experience (LE) ($\mathcal{P}_i = 0.2\frac{v_i}{u_i+v_i} + 0.8f(\text{LE})$).
- **Average approach:** The global action $a_{g_t} = d_i$ is the desired action of the human$_i$ with the median weighted effect.

- **Round Robin (RR):** The global action a_{g_t} is selected from one of the desired actions of all humans in a rotation.
- **No subgoals using 1 DQN:** The weighted effects are determined by using 1 DQN structure (Section 2.3.5.3).

We investigated multiple evaluation metrics as follows:

Fairness state results and group fairness definition To measure s_t, we need to use the satisfaction history counters by updating them per Equation 2.15. If the learning experience (LE) of the human is > 0 as measured in Equation 2.30, it will be considered satisfied. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by 36.35% and *equalized odds* fairness by 30.55% on average.** More details are shown in Figure A.5 in the Appendix.

Satisfaction performance We use $3k$ samples after FAIRO convergence($12k$ to $15k$) to examine the satisfaction value ($\frac{v}{v+u}$) histograms. FAIRO has the lowest average JSD (0.021), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO satisfaction JSD is reduced by 83.53% on average.** More numerical details are shown in Table A.1 in the Appendix.

Learning experience (LE) results We count the number of samples with positive LE and average over $3k$ samples. FAIRO achieves comparable LE results. FAIRO LE is improved by 11.4% from Average Approach, reduced by 3.4% from RR, and reduced by 3.2% from 1 DQN 3 inputs.

2.3.8 Discussion, Limitation, and Conclusion

FAIRO addresses the challenge of ensuring fairness in Human-in-the-Loop (HITL) systems by breaking it down into manageable subgoals that span a timeline for fairness-aware sequen-

tial decision-making. We introduce the FAIRO framework, which is tailored to account for the dynamic nature of human variability and preferences as they evolve over time. Our approach centers on a novel fairness state dedicated to achieving satisfaction equity, grounded in meeting human preferences. This concept of fairness is designed to enhance CPS that involve human interactions by ensuring equitable satisfaction levels. Nevertheless, our current model of fairness presents a limitation as it does not fully consider how interpersonal interactions may influence individuals' perceptions of satisfaction, indicating an area for future enhancement.

2.4 Discussion

FinA and FAIRO both take fairness in an adaptive system to mean keeping people's accumulated experience close to one another over time, as opposed to maximizing any one person's satisfaction or the group's average. What they establish jointly, beyond the results reported in their own sections, is that this goal can be implemented in two ways, and that choosing between them is a trade-off. FinA makes the objective explicit and interpretable: fairness is the minimization of a psychologically grounded adverse effect, solved by optimization at each step. That yields an auditable definition of what "fair" means, but requires the problem to be re-solved from scratch whenever the accumulated state changes. FAIRO keeps the same goal but makes it learnable: by treating "raise whoever is currently worst off" as a reusable temporally abstracted sub-goal, it trades the interpretability of a closed-form objective for the scalability of a policy that improves with experience.

Both also improve fairness without cost to the task. Neither framework had to sacrifice thermal comfort, resource sufficiency, or learning progress to improve the equity of experience. That holds in the smart-home setting used for detailed study, and also in water-resource allocation and personalized learning.

Fairness here has been a property of what an adaptive system *decides*. A system trained on human data also *reveals* something about the people in the loop, and that leakage, like experience, is not shared equally. Chapter 3 asks how privacy risk is distributed, and who bears the worst of it.

Chapter 3

Fairness in Privacy for Federated Learning

3.1 Overview

Chapter 2 asked how an adaptive system should distribute *experience* fairly across the people it serves. This chapter applies that distribution-aware view to *privacy risk*. Wherever a model is trained on human data, some information about the people behind that data can be recovered from the model, and that exposure is not shared equally. The participants who are most exposed tend to be the ones who are least typical: the outliers and minorities within the population. Privacy is treated here as a distribution of risk across people, not as a single average or worst-case leakage figure. A fair system is one in which that distribution does not concentrate on the already-vulnerable.

The concrete setting is federated learning (FL), which is widely promoted as privacy-preserving because raw data never leaves a client's device. Decentralization alone does not make privacy equitable. Under the non-independent-and-identically-distributed (non-IID) heterogeneity

that characterizes real human data, the global model is driven to memorize the distinctive features of outlier clients, and memorization is what inference attacks exploit. Under an honest-but-curious server that runs a two-stage attack, first membership inference ("was this record in the training set?"), then source inference ("which client did it come from?"), the clients whose data is most unusual leak the most, and can be re-linked to their source. The burden of privacy risk thus falls hardest on those least able to blend into the crowd. Because perfect protection is unattainable under a zero-trust threat model, the goal is *privacy equity*: minimizing the disparity of privacy risk across clients without degrading model utility.

The paper presented in this chapter, **FinP** (*Fairness-in-Privacy*) [Zhao et al., 2026b], formalizes this goal and enforces it with two coordinated mechanisms. On the client side, FinP targets localized overfitting by reshaping each client's local loss: it adds a memorization penalty, implemented as a Lipschitz constraint on the local model that forces a smoother fit instead of an exact memorization of outliers, and it scales that penalty by the client's *relative* overfitting rank, so that clients who overfit more than their peers are regularized harder. That rank is measured from the geometry of learning: the curvature of the loss surface around the learned weights, read from the Hessian's top eigenvalue and trace, distinguishes a sharp, memorizing minimum from a flat, generalizing one. On the server side, the same relative ranks are reused as privacy-risk levels, and aggregation weights are set inversely to risk, so that high-risk clients exert less influence on the global model and their memorized features are not broadcast across the federation. The server's ranking then feeds back into each client's regularization in the next round, closing the loop.

Section 3.2 discusses FinP in full. Section 3.3 discusses its implications.

3.2 F*in*P: Fairness-in-Privacy in Federated Learning by Addressing Disparities in Privacy Risk

3.2.1 Introduction

The pervasive integration of machine learning (ML) into human-centric applications demands careful consideration of both its ethical implications and its privacy guarantees. Historically, algorithmic fairness research has focused primarily on preventing bias and discrimination in model decisions or predictive accuracy Kleinberg et al. [2018], Rambachan et al. [2020], Dwork et al. [2012], Kusner et al. [2017], Zhao et al. [2024], Zhao and Elmalaki [2024], Elmalaki [2021]. However, as ML systems increasingly rely on sensitive personal data, a critical yet underexplored dimension of fairness emerges: the fair distribution of privacy risks. Existing privacy definitions often focus on average or worst-case leakage across an entire dataset Mehrabi et al. [2021]. Yet, the crucial aspect of fair risk distribution—ensuring that no single individual or demographic group disproportionately shoulders the burden of privacy vulnerabilities—has received significantly less attention.

Federated Learning (FL) has emerged as an important and impactful privacy-enhancing technology to train ML models on decentralized devices. By enabling collaborative learning without centralizing raw data, FL inherently mitigates the risks of mass data collection and aligns well with stringent regulatory frameworks such as the GDPR and its Data Protection Impact Assessment (DPIA) provisions fin [2018], Chang et al. [2024], Rieke et al. [2020]. Despite these advantages, FL remains vulnerable to sophisticated privacy attacks, including Membership Inference Attacks (MIA) Shokri et al. [2017] and Source Inference Attacks (SIA) Hu et al. [2021], which aim to trace model memorization back to specific training data or participating clients. Crucially, the decentralized and statistically heterogeneous nature of FL — where data is not Independent and Identically Distributed (non-IID) — often amplifies unequal privacy leakage. When models overfit to clients with atypical or

underrepresented data, those specific clients experience significantly higher privacy risks, creating a new form of digital inequality.

Motivating Context and Societal Impact The 2024 National Public Data (NPD) breach, which exposed billions of records, powerfully underscores the societal need for fairness in privacy Staff [2024]. Although the breach was widespread, its downstream consequences disproportionately impacted vulnerable populations—such as low-income individuals and the elderly—who have fewer resources to recover from identity theft and data abuse. Acknowledging the near inevitability of some data leakage—as reflected in the "zero trust" security paradigm Rose et al. [2020]—the security community's focus must expand from solely preventing breaches to ensuring equitable risk distribution when leaks occur. Within decentralized systems such as FL, this means formally defining and mitigating disproportionate harm. Formalizing the notion of fairness-in-privacy provides policymakers with the tangible metrics required to enforce equitable protections and fosters the public trust necessary for widespread adoption of AI.

Disparate Risk in Human-Centric FL To illustrate this disparity, consider the application of FL in Human Activity Recognition (HAR) for personalized health monitoring Rieke et al. [2020], Poulain et al. [2023]. In an FL-based HAR system, each user's wearable device acts as a client Nguyen et al. [2022]. Users with physical disabilities or chronic conditions often exhibit movement patterns that deviate significantly from the "average" user in the global model. Consequently, the global model tends to memorize these unique features, making these marginalized individuals far more susceptible to re-identification via MIA Shokri et al. [2017] or SIA Hu et al. [2021]. This constitutes severe representational harm: individuals already facing societal vulnerabilities are disproportionately exposed to privacy risks simply because their data is statistically unique. Without explicitly addressing this inequitable distribution of risk, FL risks reinforcing the very social inequalities it should theoretically bypass.

Proposed Approach F*in*P addresses the critical challenge of ensuring equitable privacy protection. It is a framework designed to measure and enforce fairness in privacy within FL. To mitigate disparate privacy leakage caused by heterogeneous data and varying local training dynamics Selialia et al. [2024], we propose a novel two-pronged approach operating at both the server and the client levels:

1. **Server-Side Adaptive Aggregation:** We introduce an aggregation strategy that dynamically weights client model updates based on their estimated privacy risk. This prevents highly vulnerable, unique clients from disproportionately exposing their representations in the global model.

2. **Client-Side Collaborative Overfitting Reduction:** Complementing the server-side intervention, we tackle the root cause of vulnerability—local overfitting. By estimating each client's relative overfitting using the maximum eigenvalue of its local model's Hessian, we incorporate this metric into a local regularization term bounded by the Lipschitz constant. This encourages clients to learn generalizable representations, directly reducing their individual susceptibility to privacy attacks.

By addressing both the symptoms (unequal risk aggregation) and the root causes (local overfitting) of privacy disparity, our combined approach achieves an improvement in fairness in privacy with a negligible impact on primary task performance.

Contributions In summary, our key contributions are:

- We introduce a formal definition of fairness-in-privacy (F*in*P) tailored for real-world human-centric machine learning systems.

- We operationalize the measurement of F*in*P within Federated Learning, with a specific focus on quantifying vulnerabilities to source inference attacks (SIA).

- We propose a novel, two-pronged technical framework featuring adaptive server-side aggregation and client-side Hessian-based regularization to optimize F*in*P.

- We provide a comprehensive empirical evaluation demonstrating the efficacy of our approach using real-world dataset of a human activity recognition (HAR), and standard learning tasks (FEMNIST and CIFAR-10), confirming significant gains in fairness in privacy with minimal task utility loss.

3.2.2 Background and Related Work

Privacy in Human-Centric Systems Ensuring privacy in human-centric machine learning systems presents inherent conflicts among service utility, computational cost, and personal privacy Sztipanovits et al. [2019]. Without appropriate frameworks and incentives for secure data utilization, system deployments often face a dichotomy: policies become too restrictive, stifling innovation, or compromise private information, leading to adverse selection and a loss of public trust Jin et al. [2016], Mulligan et al. [2016], Fox et al. [2021], Goldfarb and Tucker [2012]. Consequently, extensive research has focused on establishing privacy-preserving techniques and auditing platforms for human-in-the-loop and decision-making systems Abadi et al. [2016], Cummings et al. [2019], Taherisadr et al. [2023], Taherisadr and Elmalaki [2024], Jagielski et al. [2020], Raji et al. [2020], Elmalaki et al. [2022]. However, recognizing the "zero trust" reality that perfect privacy is often mathematically or practically unattainable, FinP examines privacy through an equity lens. We investigate how to ensure a fair distribution of harm when privacy leakage inevitably occurs, bridging the technical challenges of ML with the ethical imperatives of equitable protection.

Privacy Disparity and Inference Attacks in Federated Learning. Federated Learning (FL) enables the collaborative training of models without centralizing raw data McMahan et al. [2017], making it highly suitable for privacy-sensitive domains navigating regulations, such as GDPR Nguyen et al. [2022], Zhang et al. [2021]. Despite its decentralized architec-

ture, FL introduces new attack surfaces. The statistical heterogeneity inherent in FL does not only affect model performance — it systematically amplifies the disparity in privacy risk between participating clients. When client data distributions diverge significantly, models tend to memorize the unique features of outlier clients, making those clients disproportionately vulnerable to inference attacks. Foundational work on disparate vulnerability Kulynych et al. [2022] has demonstrated that privacy leakage is rarely uniform; models disproportionately memorize and expose the data of underrepresented subgroups. Membership Inference Attacks (MIA) exploit this memorization by determining whether a specific data record was used during training, effectively probing the model's generalization boundary Shokri et al. [2017], Hu et al. [2022]. Although MIA captures global membership leakage, it does not reveal which client contributed to the targeted record Gu et al. [2022], Zhu et al. [2024]. Source Inference Attacks (SIA) extend this threat by identifying the exact originating client of a training record, directly exploiting localized overfitting caused by non-IID data distributions Hu et al. [2021]. Despite pseudonymity, SIA poses a serious risk because mapping sensitive data to a distinct client node links it to a real-world context such as a device or location, effectively breaking anonymity. Beyond membership-based attacks, Property Inference Attacks (PIA) represent a distinct threat class that targets macro-level statistical properties of a client's local dataset — such as demographic distributions or class ratios — rather than individual records Melis et al. [2019]. Unlike SIA and MIA, which are rooted in localized memorization and overfitting, PIA exploit the structural properties embedded in gradient updates during the aggregation process and, therefore, operate from a fundamentally different threat vantage point.

Crucially, the inherent statistical heterogeneity (Non-Independent and Identically Distributed (non-IID) data) of human-centric FL amplifies these vulnerabilities. When data distributions differ widely among clients, individual model updates become highly distinct. Malicious actors can exploit these distributional differences to trace memorized features to specific clients Zhu et al. [2021]. This distinctiveness is especially pronounced in datasets like Human

Activity Recognition (HAR), where unique physical movements or conditions make outliers highly susceptible to SIAs, creating an unequal landscape of privacy risk.

Existing defenses such as Differential Privacy via DP-SGD Abadi et al. [2016] address these risks by injecting uniform noise into gradient updates. However, as demonstrated in FinP, geometry-blind noise injection degrades global utility without resolving the disparity in per-client privacy exposure; in fact, it can exacerbate unfairness by failing to mask the highly distinct updates of severely skewed clients while unnecessarily penalizing well-represented ones.

From Utility Fairness to Privacy Fairness. Fairness-aware federated learning approaches have emerged to address performance inequity between clients, including techniques such as FedAvg Hu et al. [2023] with client reweighting, FairFed Ezzeldin et al. [2023], and personalized federated learning methods Tan et al. [2022], Chen et al. [2022], as well as recent work exploring the tension between fairness interventions and privacy leakage Bendoukha et al. [2025]. Recent work also investigated the use of FL for pluralistic alignment of Large Language Models with a fairness lens of aggregation of human preference Srewa et al. [2026, 2025a,b].

However, these approaches primarily frame fairness in terms of equitable model accuracy or predictive benefit across client groups, and do not explicitly formalize or optimize for the equitable distribution of privacy risk. The critical dimension of client-level privacy equity — ensuring that no individual client disproportionately bears the burden of privacy vulnerability due to the statistical uniqueness of their data — remains largely unaddressed in the prior literature. F*in*P uniquely targets this gap by formalizing fairness-in-privacy as a first-class objective, directly mitigating memorization-based privacy disparities caused by non-IID data distributions **exploited by SIA**, and distinguishing itself from existing defenses that focus either on global privacy guarantees or on utility fairness without addressing the structural

inequity of per-client privacy exposure.

3.2.3 Threat Model and Problem Formulation

Although Federated Learning (FL) mitigates the risks of centralized data storage, it remains vulnerable to inference attacks launched by the aggregating server.

Assumptions and Scope F*in*P operates under two core assumptions. First, the server is "honest-but-curious": it faithfully executes the FL protocol but actively attempts to infer sensitive client information from received model updates and curvature metrics. Second, clients are honest and cooperative: they follow the protocol and submit genuine updates without manipulation. Under these assumptions, F*in*P targets inference attacks driven by localized overfitting and memorization in non-IID settings — specifically Source Inference Attacks (SIA), which identify the originating client. SIA is particularly risky because mapping a sensitive record to a specific client node links it to a real-world context such as a device or location, breaking pseudonymity through contextual profiling. F*in*P suppresses this memorization by flattening the loss landscape, reducing the attacker's confidence. F*in*P does not address malicious server behavior, manipulative clients, gradient reconstruction attacks, or Property Inference Attacks, as these operate outside the memorization-based threat model.

Threat Model A primary privacy threat in this decentralized setting is a two-stage inference attack executed by the server: a Membership Inference Attack (MIA) followed by a Source Inference Attack (SIA).

- **Membership Inference Attack (MIA):** The server first attempts to determine if a specific target data point x was utilized in the training set D_{θ_g} of the global model θ_g. This is formally defined as the probability that x belongs to the training data:

$$MIA(\theta_g, x) = P(x \in D_{\theta_g})$$

- **Source Inference Attack (SIA):** If the MIA successfully indicates that x was part of the training corpus, the server proceeds to identify the origin of the data. The server analyzes the updates of the local model θ_i to determine the probability that a specific client C_i is the source of the record x:

$$SIA(\theta_i, x) = P(C_i \mid x, \theta_i)$$

As demonstrated in the prior literature Hu et al. [2021], the concatenation of these attacks compromises the anonymity of the client. Furthermore, auditing and completely preventing MIA and SIA within FL presents inherent mathematical and practical limitations Chang et al. [2024].

Problem Formulation: Disparate Privacy Risk Rather than attempting the mathematically fraught task of completely eliminating SIA, our work focuses on the resulting *disparity* in privacy risk across the client federation. In heterogeneous (non-IID) FL environments, clients with unique or outlier data distributions are highly prone to local overfitting during training. This overfitting causes their local model updates (θ_i) to disproportionately memorize their unique data points, rendering them significantly more vulnerable to SIAs than clients with more "average" data. To exploit this vulnerability, The attacker calculates the prediction loss of the target record across every individual client's model on the server. The reason is that the local model belonging to the true source will naturally yield the lowest prediction loss. The attacker therefore simply identifies the client with the minimum loss as the most probable source of the target record. This creates a systemic inequity in privacy protection.

Given this threat model, our objective is to mitigate the disparate impact of SIAs by ensuring an equitable distribution of the inherent privacy risks between all clients. To achieve this F*in*P, our framework pursues two core objectives:

(**O1**): **Addressing the Symptoms (Server-Side):** Develop a dynamic aggregation method-

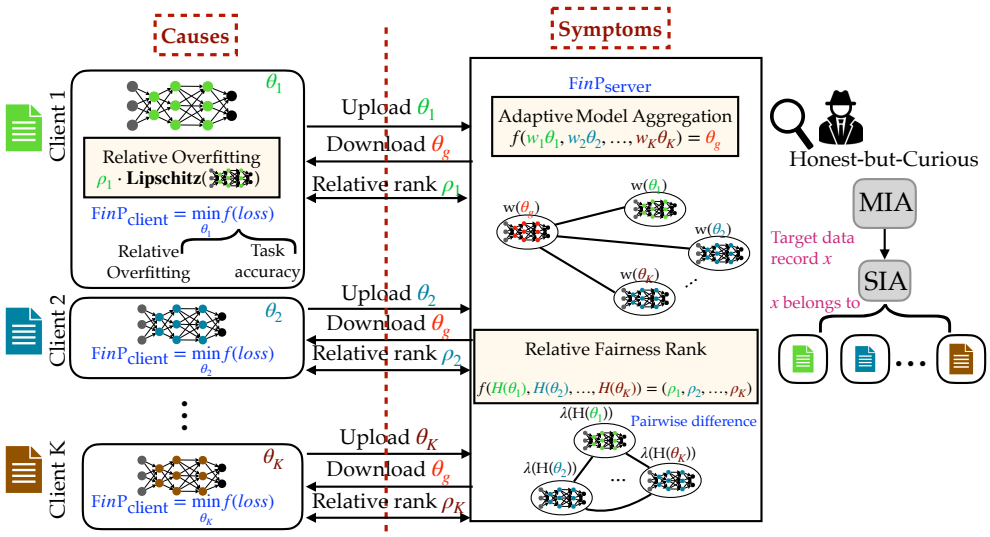


Figure 3.1: Overview of the F*in*P framework. The system achieves fairness-in-privacy by establishing a feedback loop that simultaneously addresses the symptoms of privacy disparity (via server-side adaptive aggregation) and the root causes (via client-side collaborative overfitting reduction).

ology on the server that quantifies client vulnerability and adjusts their contributions to the global model, ensuring a fair distribution of privacy risk without degrading global utility.

(**O2**): **Addressing the Root Causes (Client-Side):** Provide actionable feedback mechanisms to highly vulnerable clients, enabling them to apply targeted local regularization. This empowers clients to reduce their local overfitting, thereby lowering their individual susceptibility to SIAs and improving the overall fairness in privacy of the system.

3.2.4 The F*in*P Framework in Federated Learning

The core objective of the F*in*P framework is to ensure that privacy risks are equitably distributed among all participating clients, preventing any single client from bearing a disproportionate burden of vulnerability. In practical federated learning (FL) deployments, the profound heterogeneity of client data distributions, varying computational resources, and distinct local training dynamics naturally lead to stark disparities in privacy risk. Traditional privacy defenses that merely attempt to lower the *average* risk across the federation

are insufficient for human-centric systems; effective protection demands that we proactively prevent disproportionate risk burdens on minority or outlier clients.

To achieve true fairness-in-privacy, F*in*P operationalizes a holistic two-pronged approach, as illustrated in Figure 3.1. We argue that systemic privacy equity requires simultaneous intervention at both the server level (during model aggregation) and the client level (during local training).

Addressing the Symptoms (Server-Side Intervention) Server-side interventions are critical for managing the immediate impacts of existing privacy disparities. F*in*P introduces an adaptive aggregation mechanism that dynamically scales the weights of client updates based on their estimated privacy risk. By systematically reducing the aggregation weight of highly vulnerable clients, we prevent their overfit, memorized representations from unduly influencing the global model. This limits the exposure of their unique data to the rest of the federation and restricts an adversary's ability to execute Source Inference Attacks (SIAs). However, while adaptive aggregation effectively mitigates the immediate symptoms of privacy unfairness, it does not cure the underlying vulnerability of clients themselves.

Addressing the Root Causes (Client-Side Intervention) To dismantle the root cause of privacy disparity, we must look to local training dynamics—specifically, local overfitting. Extensive prior literature has established the theoretical and empirical connection between model overfitting and increased privacy risks Hu et al. [2021], Yeom et al. [2018], Shokri et al. [2017]. Because overfit models memorize specific properties of their training data rather than learning generalized features, they serve as the primary vulnerability exploited by MIAs and SIAs. Consequently, when a client's local model heavily overfits its non-IID data, that client's susceptibility to SIAs spikes, creating the very disparity we aim to eliminate.

To counteract this, F*in*P introduces a **collaborative overfitting reduction strategy** at the client level. This proactive mechanism ranks clients based on an estimation of their

relative overfitting compared to the federation. By incorporating this server-provided rank into a customized local regularization scheme, F*in*P forces vulnerable clients to learn more generalizable representations. This actively suppresses the memorization of outlier data, shrinking the initial disparity in privacy risk *before* the models are ever transmitted to the server.

The Closed-Loop System Ultimately, F*in*P functions as a tightly coupled closed-loop system. The server utilizes incoming client models to calculate adaptive aggregation weights and relative overfitting ranks; in turn, the clients utilize this targeted feedback from the server to regularize their subsequent local training rounds. By simultaneously addressing both the symptoms and the root causes of privacy disparity, F*in*P establishes a structurally fair FL environment.

We formally define the mathematical mechanics of the client component in Section 3.2.4.1 and the server component in Section 3.2.4.2.

3.2.4.1 Addressing the Root Causes: Client-Side Collaborative Regularization

Although server-side aggregation mitigates the immediate symptoms of privacy unfairness, achieving structural fairness-in-privacy (F*in*P) requires dismantling the root cause: local overfitting. In a heterogeneous FL system with K clients, non-IID data create distributional shifts that cause certain outlier clients to overfit more than others. Because overfitting directly correlates with the memorization of training data, these specific clients become disproportionately susceptible to Source Inference Attacks (SIAs). Our goal is to enforce an equitable privacy risk by proactively reducing this localized memorization.

To achieve this, we introduce a collaborative, feedback-driven regularization strategy. Recent literature demonstrates that the top Hessian eigenvalue ($\lambda_{\max}$) and the Hessian trace (H_T) are powerful metrics to characterize the loss landscape and the generalization capabilities

of neural networks Jiang et al. [2020]. A sharp loss landscape (indicated by high $\lambda_{\max}$ and H_T) implies severe overfitting and a high vulnerability to weight perturbations, while lower values indicate smoother, more generalized and thus more privacy-preserving local models.

Because we are specifically interested in F*in*P—which focuses on *disparity* rather than absolute risk—we quantify how much more vulnerable a client is compared to its peers. We define each client's relative overfitting by calculating the average pairwise difference across the top Hessian eigenvalue ($\lambda^k_{\max}$) and Hessian trace (H^k_T) of their local model θ_k against all other j clients:

$$\bar{\Delta}_k = \frac{1}{K-1} \sum_{j=1, j\neq k}^{K} |\lambda^k_{\max} - \lambda^j_{\max}| \tag{3.1}$$

$$\bar{H}_k = \frac{1}{K-1} \sum_{j=1, j\neq k}^{K} |H^k_T - H^j_T| \tag{3.2}$$

These metrics are then normalized and combined to compute the client's *Overfitting Relative Rank* (ρ_k), which serves as a proxy for their relative privacy risk disparity:

$$\rho_k = \frac{1}{2} \left(\frac{\bar{\Delta}_k}{\max(\bar{\Delta})} + \frac{\bar{H}_k}{\max(\bar{H})} \right) \tag{3.3}$$

Clients compute their own top Hessian eigenvalue and Hessian trace locally during training. The server only computes the relative rank ρ_k after receiving these scalar metrics. The server then transmits this scalar rank ρ_k to client k as a targeted feedback.

Upon receiving ρ_k, the client incorporates this overfitting rank into their subsequent local training rounds using a dynamic regularization term. We utilize the Lipschitz constant, approximated by the spectral norm of the Jacobian matrix ($||J_k||$) Liu et al. [2020]. Constraining the Lipschitz constant forces the model to learn smoother functions, thereby resisting the

memorization of outlier data points. The modified local loss function for client k becomes:

$$\mathcal{L}_k^* = \mathcal{L}_k + \beta \cdot \rho_k \cdot ||J_k|| \tag{3.4}$$

$$\mathrm{F}in\mathrm{P}_{\text{client}} = \min_{\theta_k} \mathcal{L}_k^* \tag{3.5}$$

Here, $\mathcal{L}_k$ is the original local loss function, $||J_k||$ is the Jacobian penalty, and β is a task-dependent parameter that adjusts the baseline impact of Lipschitz regularization. The critical innovation is the inclusion of ρ_k, which acts as an **adaptive** multiplier. In each communication round, clients with a higher relative overfitting rank are subjected to mathematically stronger regularization. This collaborative, penalty-based approach effectively throttles the most vulnerable clients, forcing them to generalize, and thereby actively shrinking the privacy risk disparity across the federation.

Computational Practicality and Privacy Guarantee Computing and transmitting the exact Hessian matrix is computationally prohibitive for deep neural networks and violates the strict data-minimization guarantee of FL. To ensure that our framework remains scalable, clients do not compute the full Hessian; instead, they efficiently estimate only the top eigenvalue and the trace.

To minimize local computational overhead on edge devices — including battery-constrained mobile phones and resource-limited edge nodes — clients employ efficient Hessian-vector product (HVP) methods, specifically power iteration for the top eigenvalue and Hutchinson's method for the trace, entirely avoiding full Hessian instantiation. Critically, because FinP's fairness objective relies solely on the relative ranking of client vulnerability rather than the absolute values of eigenvalues, these lightweight approximations are fully sufficient to achieve privacy equity without requiring exact Hessian computation. This design choice substantially reduces computational cost per-round, making FinP practical for deployment on resource-constrained hardware typical in the real-world federated participants.

F*in*P does not require additional communication rounds to transmit these metrics. The scalar sharpness values are simply appended to and transmitted concurrently with the standard local model weights during existing federated communication rounds, incurring only negligible additional bandwidth overhead and fully preserving the communication efficiency of standard federated learning.

Furthermore, while F*in*P requires clients to transmit these local curvature metrics with their model updates(θ_k), this does not violate the FL privacy principles. These metrics are one-dimensional scalars representing macroscopic geometric properties of the local loss landscape. Because mapping a local dataset to these two aggregate scalars is a heavily non-injective and non-invertible transformation, transmitting them provides negligible mutual information regarding the raw training data. It is mathematically intractable for an honest-but-curious server to reconstruct raw local data or execute gradient-inversion attacks relying solely on these curvature scalars.

3.2.4.2 Addressing the Symptoms: Server-Side Adaptive Aggregation

The core symptom of privacy unfairness in FL is the pronounced disparity in vulnerability to Source Inference Attacks (SIAs) across the client federation. While client-side regularization (detailed in Section 3.2.4.1) addresses the root cause of local overfitting, the server must simultaneously manage the observable symptoms present in the incoming model updates. Our objective is to design a server-side aggregation strategy that actively quantifies this privacy risk and adjusts the global model update to enforce an equitable distribution of vulnerability. We explicitly define this objective as minimizing the variance of the privacy risk distribution:

$$
\begin{aligned}
\mathrm{F}in\mathrm{P}_{\text{server}} &= \min_{\mathbf{w}\in\mathcal{W}} \text{Disparity}(\mathbf{p}(\mathbf{w})) \\
&= \min_{\mathbf{w}\in\mathcal{W}} \left\| \mathbf{p}(\mathbf{w}) - \frac{1}{K}\mathbb{1}^T\mathbf{p}(\mathbf{w}) \otimes \mathbb{1} \right\| + \left\| \frac{1}{K}\mathbb{1}^T\mathbf{p}(\mathbf{w}) \right\|
\end{aligned}
\tag{3.6}
$$

Here, $\mathbf{p}(\mathbf{w})) = [p_1(\mathbf{w})), \ldots, p_K(\mathbf{w}))]$ is the vector of privacy risk indicators across the K clients given the aggregation weights $\mathbf{w}$, $\mathbb{1}$ is a vector of ones of length K, and $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^K \mid \sum_{k=1}^{K} w_k = 1, w_k \geq 0 \ \forall k\}$ is the set of valid aggregation weights.

To satisfy the $\mathrm{F}in\mathrm{P}_{\text{server}}$ objective, the server must aggregate local models in a manner that minimizes this disparity metric without severely degrading global utility. We evaluate two strategies for achieving this:

1. **Strategy 1: Adaptive Lightweight Aggregation (ALA) (Proposed Solution)**
   Our primary method acts as a highly efficient, heuristic solution to the disparity minimization objective, specifically designed to be scalable for real-world, resource-constrained environments.

   - **Risk Symptom Quantification (p_k):** The server re-utilizes the normalized Overfitting Relative Rank (ρ_k) described in Equation 3.3, as a proxy for the privacy risk symptom ($p_k \propto \rho_k$). A higher ρ_k signifies severe local overfitting and greater vulnerability to SIAs.

   - **The Solution Mechanism:** The ALA technique achieves the disparity minimization objective by dynamically calculating the specific aggregation weights $\alpha_k \in \mathcal{W}$ to be inversely proportional to the risk symptom ρ_k. High-risk clients are explicitly granted less influence over the global model, proactively driving the global model toward a state where the risk vector $\mathbf{p}$ is balanced:

$$\theta_{global} \leftarrow \sum_{k=1}^{K} \alpha_k \theta_k, \quad \text{where} \quad \alpha_k = \frac{1-\rho_k}{\sum_{i=1}^{K}(1-\rho_i)}, \tag{3.7}$$

where θ_{global} is the global model parameters and θ_k is the local model parameters for client k. This inverse weighting mechanism provides the exact control necessary to solve Equation 3.6. By actively restricting the impact of the most vulnerable clients, ALA prevents their memorized, outlier data from being embedded into the global model. This enforces an equitable distribution of privacy risk without introducing significant computational overhead.

2. **Strategy 2: Server-Side Risk Quantification (PCA-based Baseline)**

   For comparison only, a traditional alternative for quantifying the symptom of risk p_k is computing the Principal Component Analysis (PCA) distance between each client's model update and the global average update Durmus et al. [2021]. This distance acts as a purely server-side proxy for outlier behavior.

   If this PCA distance (PCA_d) is used as the risk vector $\mathbf{p}(\mathbf{w})$, the aggregation process would similarly solve the $\text{FinP}_{\text{server}}$ disparity minimization objective by adjusting weights based on these distances. However, while formally sound, computing the PCA distance for the high-dimensional model updates typical in modern human-centric FL is computationally expensive Candès et al. [2011]. The resulting latency makes it highly impractical for real-world, large-scale FL deployments, reinforcing that the lightweight, Hessian-proxy ALA strategy is the superior and preferred mitigation method for our framework.

3.2.5 Evaluation

3.2.5.1 Federated Learning System Setup

To demonstrate the real-world applicability of F*in*P to mitigate privacy disparities, we evaluated our framework in two distinct federated learning deployments. We specifically selected

datasets that reflect the highly heterogeneous, non-IID nature of practical edge-computing and mobile sensing environments, where outlier clients are naturally prone to local overfitting and heightened vulnerability to Source Inference Attacks (SIAs). In particular, we structured our evaluation around two primary case studies: Human Activity Recognition (HAR) as a human-centric, privacy-sensitive application, and CIFAR-10 and FEMNIST as a standard benchmark for controlled, heterogeneous visual data.

Baselines and Ablations To rigorously isolate the contributions of our framework, we compare the three standard and state-of-the-art baselines in various configurations as follows:

• **HAR dataset:** we evaluated five configurations: (1) Baseline FL: Standard FedAvg aggregation, adapted from prior SIA literature Hu et al. [2023]. (2) F*in*P $_{\text{server}}$: Isolating our adaptive server-side aggregation without client collaboration (Equation 3.6). (3) F*in*P $_{\text{client}}$: Isolating our collaborative client-side regularization without adaptive server aggregation. (4) F*in*P-Full: The complete proposed closed-loop framework. (5) Differential Privacy (DP-SGD) Abadi et al. [2016]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$.

• **CIFAR-10 dataset:** we evaluated: (1) Baseline FL: Standard FedAvg from Hu et al. [2023]. (2) FedAlign: A state-of-the-art FL method specifically designed to address non-IID data heterogeneity [Mendieta et al., 2022]. (3) F*in*P (ResNet56): Our framework applied to the same ResNet architecture used by FedAlign for a direct performance comparison. (4) F*in*P Ablation (CNN): Our framework applied to a standard CNN to evaluate the sensitivity of the impact factor $\beta \in \{0.05, 0.1, 0.3, 0.5\}$, as described in Equation 3.4. (5) Differential Privacy (DP-SGD) Abadi et al. [2016]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$. (6) Additionally, we evaluate F*in*P using both the Adaptive Lightweight Aggregation (ALA) and the PCA-based server-side strategies to validate computational practicality.

• **FEMNIST dataset:** we evaluate: (1) Baseline FL: Standard FedAvg from Hu et al. [2023]. (2) F*in*P: Our framework with the impact factor $\beta \in \{0.5, 0.75, 1\}$, as described in Equation 3.4. (3) Differential Privacy (DP-SGD) Abadi et al. [2016]: Applying DP-SGD against SIA attack with noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$. (4) MIA attack: Launching White-Box Membership Inference Attack (MIA) to evaluate F*in*P performance vs. Baseline (FedAvg) and DP-SGD.

Setup for HAR The UCI HAR Dataset Reyes-Ortiz et al. [2013] serves as our real-world application. In modern ubiquitous computing and mobile sensing environments, continuous sensor data (such as accelerometer and gyroscope readings) is routinely collected from users' personal smartphones or wearable devices. Because physical behaviors, daily routines, and gaits are unique to an individual, HAR data is inherently non-IID and deeply sensitive [Concone et al., 2022, Tu et al., 2021]. Preserving the natural user-based partition ensures the authentic structure of the data is maintained; models trained on users with distinct or outlier behaviors are prime targets for SIAs.

The HAR dataset comprises data from 30 subjects (aged 19–48), treated as $K = 30$ individual FL clients, performing six daily activities. We allocated 70% of each client's data for training (using 5-fold cross-validation) and 30% for independent testing. All data was preprocessed with noise filters and segmented using a 2.56-second sliding window with a 50% overlap. We utilized a Temporal Convolutional Network (TCN) Bai et al. [2018] for the architecture of the local models, which captures temporal dependencies via causal convolutions. We trained over 20 global communication rounds using FedAvg. Clients trained locally with a batch size of 64, a learning rate of 0.001 (Adam optimizer), 1 local epoch per round, and an impact factor $\beta = 2$. More details on the HAR dataset are in Appendix B.1.

To establish a rigorous Differential Privacy (DP) baseline, we implemented Record-Level DP-SGD Abadi et al. [2016] utilizing the Opacus library. For all DP-enabled experiments, we

transitioned from the Adam optimizer to Stochastic Gradient Descent (SGD) with a learning rate of 0.01, while keeping all other federated configurations (batch size, local epochs, and communication rounds, etc) identical to the standard baseline. A critical challenge in DP-SGD is selecting an optimal clipping threshold (C) that aggressively bounds the sensitivity of extreme outliers while preserving the underlying gradient signal. To ensure an empirical and fair calibration, we executed a non-private warm-up run of the global model for a single epoch and recorded the unclipped L_2 gradient norms across all client batches. We fixed the clipping threshold at C=5, which closely approximates the 90th percentile of the unclipped gradients. By adopting this empirically derived threshold, we formally bounded the model's sensitivity while maintaining optimal learning momentum. To comprehensively evaluate the resulting privacy-utility trade-off, we swept the Gaussian noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$, corresponding to low, medium, and high privacy budgets, respectively.

Setup for CIFAR-10 To stress-test F*in*P under mathematically controlled degrees of data imbalance—simulating visual heterogeneity across a network of personal devices or edge cameras—we utilized the CIFAR-10 dataset [Krizhevsky et al., 2009]. We partitioned the data among $K = 10$ clients using a Dirichlet distribution $Dir(\alpha)$ with $\alpha = 0.5$, a standard approach in SIA literature for simulating unbalanced subsets [Mendieta et al., 2022, Hu et al., 2021]. Figure 3.2 visualizes the severe data skew among clients when tested at $\alpha = 0.1$. We employed ResNet56 [He et al., 2016] for the primary evaluation to match the FedAlign baseline, and a standard CNN Hu et al. [2023] for the β ablation studies. Similar to the HAR setup, training occurred over 20 global rounds. The default impact factor for the client-side regularization was set to $\beta = 0.3$. We deployed DP-SGD with CNN and same configuration. Clipping threshold is fixed at $C = 1.75$ ($\approx$ the 90th percentile of the unclipped gradients) and noise multiplier $\sigma_{DP} \in \{0.5, 1, 2\}$ for low, medium, high privacy budgets. More details on the CIFAR dataset are in Appendix B.1.

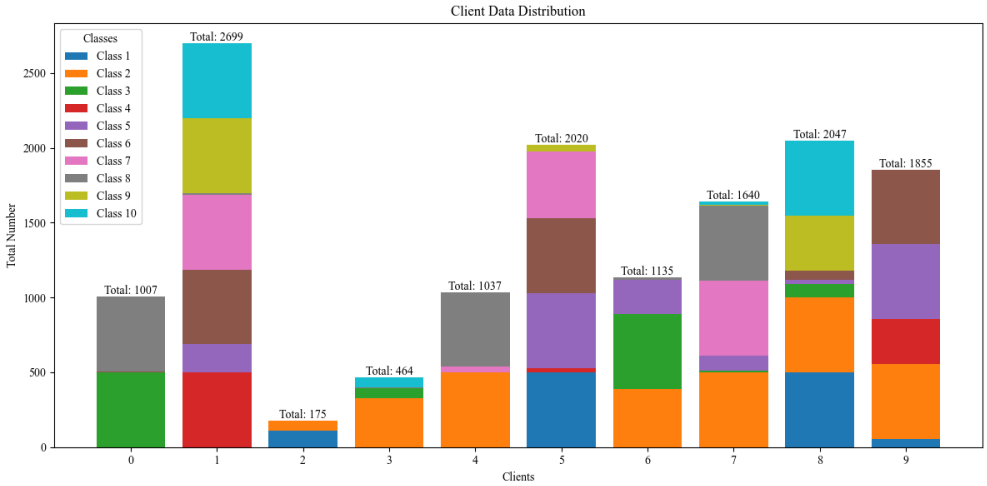


Figure 3.2: CIFAR dataset profile demonstrating severe non-IID data distribution among clients after Dirichlet sampling ($\alpha = 0.1$).

Setup for FEMNIST We evaluate on FEMNIST Dataset additionally, a handwritten character recognition benchmark with 62 classes and natural non-IID structure induced by writer identity. We simulate 10 federated clients by sampling 10 distinct writers; each client owns only that writer's examples. We employ a lightweight CNN as the training model. More details are in Appendix B.1.

Empirical Evaluation Setup To simulate this attack pipeline for our empirical evaluation, we randomly sampled training data from each client's local dataset and pooled them into a unified dataset of target records. By launching the SIA against these known records and calculating the attack success rate, we can directly measure both the absolute privacy leakage and, crucially, the *disparity* of this vulnerability across the federation.

Hardware Profile All experiments were conducted on a single NVIDIA A30 GPU with 24 GB of memory, simulating the central aggregating server's computational environment.

3.2.5.2 Metrics for Comparison

To comprehensively evaluate the effectiveness of F*in*P in achieving fairness in privacy, we employ the following key metrics encompassing privacy fairness, absolute privacy, utility,

and efficiency.

1. Privacy Fairness Metrics (Disparity Quantification) To quantify "Fairness in Privacy," we must measure the dispersion of Source Inference Attack (SIA) vulnerability across the K participating clients. We introduce three metrics to capture both the accuracy and the confidence of the adversary's attack.

- **Disparity in SIA Accuracy (CoV & FI):** To formalize the notion of privacy fairness, we draw on the concept of group fairness as established in literature Jain et al. [1984]. Applying the Coefficient of Variation (CoV) to client-level privacy risk treats each client as a protected group, where a low CoV indicates an equitable distribution of privacy exposure across the federation. Unlike minimax or worst-case approaches that optimize solely for the single most vulnerable client, minimizing CoV ensures system-wide privacy equity, preventing minority clients from suffering disproportionately high privacy vulnerability while preserving equitable protection across the entire federation.

  For a federation of K clients, the empirical SIA accuracy for a specific client i is calculated as:

$$\text{SIA}_{acc_i} = \frac{\text{Number of correct SIA identifications for client } i}{\text{Total target records attributed to client } i} \tag{3.8}$$

  Let $\mu = \frac{1}{K}\sum_{i=1}^{K} \text{SIA}_{acc_i}$ be the mean SIA accuracy across the federation, and σ be the standard deviation. The CoV is computed as:

$$\text{CoV}(\text{SIA}_{acc}) = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{K}\sum_{i=1}^{K}(\text{SIA}_{acc_i} - \mu)^2}}{\mu} \tag{3.9}$$

  To map this variance to an intuitive fairness percentage between 0 and 1 (where 1 represents perfect equity), we apply the Fairness Index (FI) transformation:

$$\text{FI}(\text{SIA}_{acc}) = \frac{1}{1 + \text{CoV}(\text{SIA}_{acc})^2} \tag{3.10}$$

  A higher FI value indicates that the vulnerability is equitably distributed, preventing any single user from becoming an extreme outlier target.

- **Equal Opportunity Difference in Privacy (EOD):** Because SIA_{acc_i} represents the empirical True Positive Rate (TPR) of the adversary correctly identifying client i, we can map traditional algorithmic group fairness metrics—specifically Equal Opportunity—directly to privacy [Hardt et al., 2016]. By treating individual clients as protected groups, Equal Opportunity mandates that the TPR of the adversary's attack should be uniform across the federation. We quantify the maximum violation of this principle using the Equal Opportunity Difference (EOD):

$$EOD = \max_{i,j} |\text{SIA}_{acc_i} - \text{SIA}_{acc_j}| \quad \forall i, j \in \{1, \ldots, K\} \tag{3.11}$$

  A lower EOD demonstrates that the adversary cannot exploit specific outlier users more easily than the average user.

- **Disparity in Attack Confidence (Loss CoV):** While SIA accuracy measures discrete attack success, a rigorous privacy evaluation must also address the adversary's confidence. SIAs exploit the correlation between localized memorization and low target-record prediction loss. Clients with highly overfitted local models will exhibit significantly lower prediction losses on their target records, granting the adversary high statistical confidence. F*in*P actively aims to flatten this inter-client loss landscape. We quantify this mitigation using the CoV and FI of the client prediction losses on target records, denoted as CoV(Loss) and FI(Loss), defined similarly to Equations 3.9 and 3.10.

2. Absolute Privacy Metrics Achieving fairness by artificially degrading the privacy of secure clients to match the most vulnerable ones is a failure mode in privacy engineering. To verify that F*in*P enforces equitable privacy without increasing overall leakage, we track two global metrics:

- **Mean(SIA_{acc}):** The average attack success rate across all clients and communication rounds.

- **Max(SIA_{acc}):** The absolute highest attack success rate observed for any single client.

Table 3.1: Results of HAR using the two approaches (ALA and PCA) of server aggregation in comparison to the Baseline Hu et al. [2023] and DP-SGD Abadi et al. [2016].

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA_{acc}) ↓	Max (SIA_{acc})↓	CoV(SIA_{acc})/FI(SIA_{acc})	CoV(loss)/FI(Loss)	EOD ↓	Conv.round
Baseline Hu et al. [2023]	74.10	76.97	19.34	22.40	0.94/0.54	0.244/0.938	0.48	9
F*in*P $_{\text{client}}$	73.39	75.86	18.87	23.30	0.97/0.53	0.237/0.941	0.49	9
F*in*P $_{\text{server}}$ (PCA)	**75.01**	**77.84**	**18.57**	**21.60**	0.99/0.52	0.246/0.937	0.54	11
F*in*P (PCA)	72.73	75.22	19.70	23.50	**0.89/0.57**	0.235/0.942	0.48	10
F*in*P $_{\text{server}}$ (ALA)	71.57	73.82	19.55	26.10	0.93/0.55	**0.220/0.948**	0.55	11
F*in*P (ALA)	72.83	76.09	19.29	22.00	**0.89/0.57**	0.235/0.942	0.48	10
DP-SGD (ϵ, δ)								
(99, 10^{-5})	45.10	45.13	17.88	22.9	1.05/0.48	0.032/0.999	0.49	9
(32, 10^{-5})	41.19	42.27	16.63	18.7	1.15/0.43	0.008/1.000	0.50	9
(12, 10^{-5})	44.37	44.66	17.32	21.2	1.05/0.48	0.032/0.999	0.50	6

Lowering this maximum vulnerability is critical, as it proves we are successfully protecting the "weakest links" in the network.

3. Utility and Efficiency Metrics Finally, any deployable Privacy-Enhancing Technology must maintain the operational viability of the underlying system.

- **Global Model Utility:** We evaluate the ultimate cost of privacy by measuring the testing accuracy of the converged global model on the primary learning task.

- **System Efficiency:** We measure the impact of the F*in*P closed-loop regularization on the convergence dynamics, quantified by the total number of communication rounds required to reach convergence compared to standard baselines.

3.2.5.3 Evaluating F*in*P on the Human Activity Recognition (HAR) Dataset

We first evaluate the efficacy of F*in*P in a realistic, human-centric mobile sensing environment using the HAR dataset (More on the dataset setup is in Appendix B.1). The comprehensive results across all metrics are summarized in Table 3.1. Our findings confirm that F*in*P successfully structurally flattens the privacy risk landscape, significantly reducing vulnerability disparities across the federation while maintaining high global utility as explained below.

1. Mitigating Disparity in SIA Vulnerability (Fairness) The primary objective of F*in*P is to ensure that no single user disproportionately bears the privacy cost of participating in the federation. Our framework demonstrates a marked improvement in the equitable

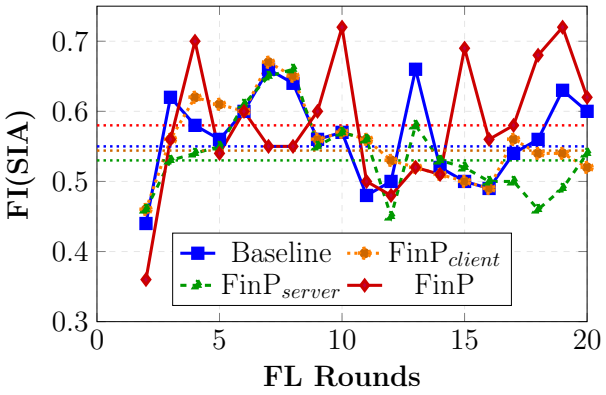


Figure 3.3: Progression of the Fairness Index (FI(SIA$_{acc}$)) over communication rounds in HAR. F*in*P (PCA) maintains a more equitable distribution of privacy risk compared to the baseline.

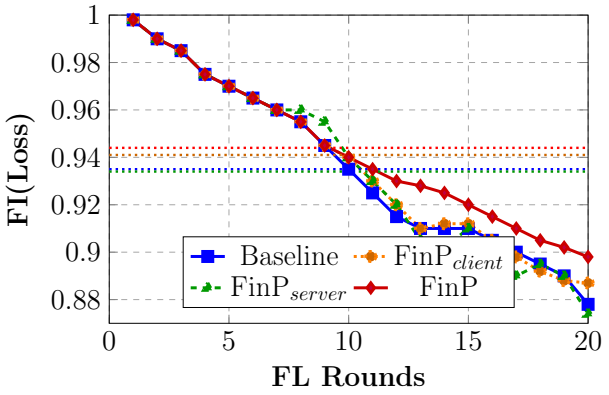


Figure 3.4: Disparity of prediction loss among clients in HAR (FI(Loss)). F*in*P (PCA) improves the inter-client loss landscape, reducing the adversary's confidence in identifying source records.

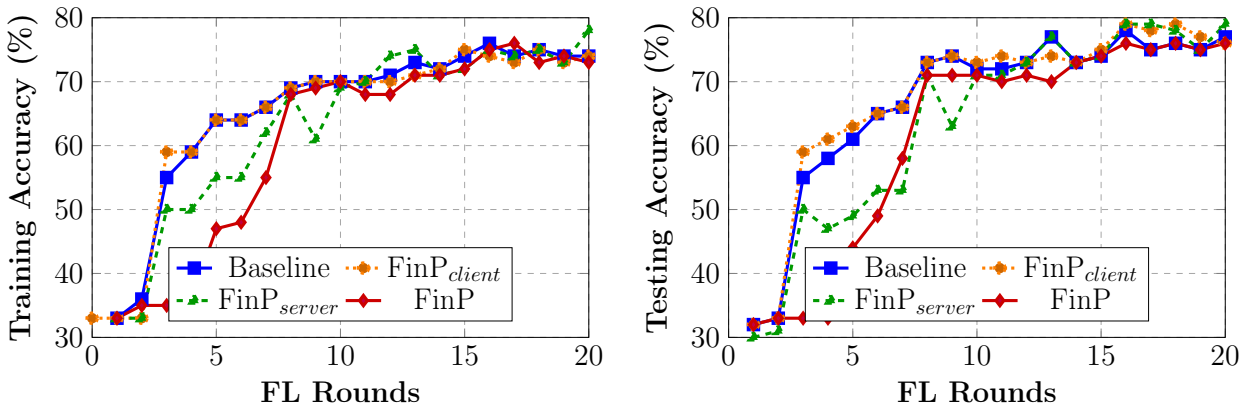
distribution of privacy risk (Figure 3.3).

- **Variance Reduction:** Compared to the Baseline FL setup [Hu et al., 2023], the full F*in*P framework reduces the Coefficient of Variation for SIA accuracy (CoV(SIA$_{acc}$)) from 0.94 to 0.89 (a 5.32% reduction). This corresponds to a 5.56% improvement in the overall Fairness Index (FI(SIA$_{acc}$)), proving that our adaptive regularization actively constrains the most vulnerable outlier models.

- **Equal Opportunity Improvement:**

  By treating individual clients as protected entities, F*in*P achieves a 5.45% reduction in the Equal Opportunity Difference (EOD), bringing it down to 0.52 (Visual illustration in Figure B.2 in Appendix B.2). This critical metric confirms that the adversary's True Positive Rate (TPR) is significantly more uniform across the network, fundamentally limiting their ability to selectively target distinct individuals based on their unique physical activities.

2. Reducing Attack Confidence (Loss Disparity) Beyond binary attack success, F*in*P effectively obscures the statistical signals adversaries rely on. By penalizing local



(a) Global model training accuracy.

(b) Global model testing accuracy.

Figure 3.5: Global model classification accuracy using HAR dataset. F*in*P (PCA) maintains competitive accuracy with minimal convergence delay.

memorization, the framework flattens the landscape of prediction losses on target records (Figure 3.4). F*in*P achieves a highly equitable FI(Loss) of 0.942 (up from the baseline's 0.938). While the absolute gain appears incremental due to the high baseline saturation, this reduction in CoV(Loss) (down to 0.235) confirms that the adversary is denied the artificially low-loss confidence signals previously emitted by overfitted outlier clients.

3. Absolute Privacy Leakage Crucially, F*in*P achieves these fairness gains without artificially degrading the privacy of secure clients. While we observe a negligible marginal increase of $< 1.1\%$ in Mean(SIA_{acc}) and $< 0.4\%$ in Max(SIA_{acc}), these fluctuations are statistically insignificant relative to the variance reduction. The network does not experience a "race to the bottom"; rather, the distribution of protection is demonstrably stabilized.

4. The Cost of Privacy: Utility and Efficiency Trade-offs A robust Privacy-Enhancing Technology must not destroy the operational value of the system. As shown in Figure 3.5, F*in*P maintains highly competitive global classification utility. The global model's testing accuracy experiences only a minimal 1.75% degradation. Furthermore, the closed-loop regularization introduces only a marginal delay in convergence, requiring ≈ 10 communication rounds to stabilize compared to the baseline's ≈ 9 rounds. This confirms that F*in*P is highly viable for real-world deployment.

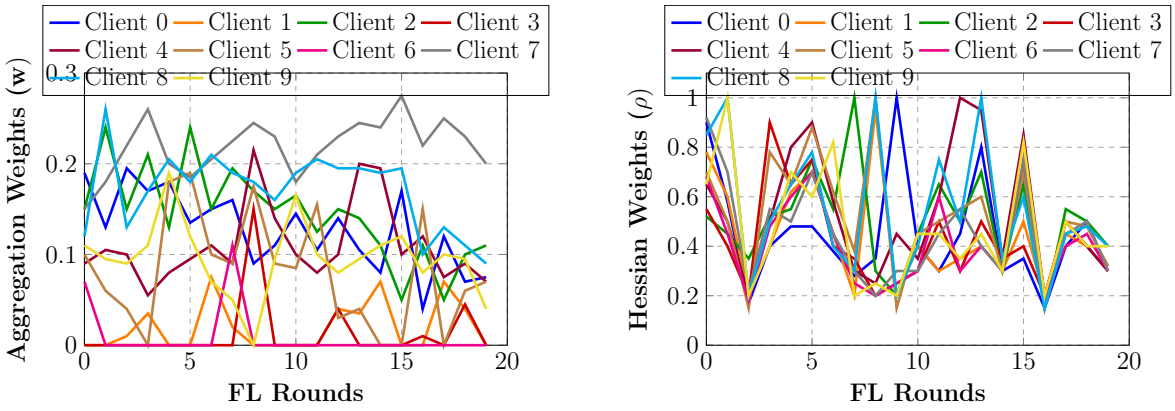
5. Component Ablation: Server vs. Client Contributions To understand the mechanics of our framework, we isolated the individual contributions of the server-side and client-side components:

- **Isolated Server Impact:** Deploying only the server-side adaptive aggregation (FinP$_{\text{server}}$ with PCA) revealed nuanced behavior. While it successfully reduced the variation in global model distance (PCA$_d$) by 1.3% and marginally lowered absolute SIA metrics, it caused a slight negative shift in the fairness indices (FI(SIA$_{acc}$) dropped by 0.02). This proves that purely server-side, reactive re-weighting addresses the symptom of model divergence but is insufficient to enforce structural fairness in privacy.

- **The Synergy of the Closed Loop:** Combining the server aggregation with the client-side collaborative regularization (FinP$_{\text{client}}$) yielded the highest fairness gains. The server-side re-weighting (Equation 3.6) acts synergistically with the client-side adaptive regularizer (ρ_k, Equation 3.5). Figure 3.6 visualizes this dynamic interplay over time.

- **Computational Practicality of ALA:** When replacing the expensive PCA computation with our proposed Adaptive Lightweight Aggregation (ALA) proxy, FinP achieved identical fairness improvements while consuming only 17% of the computation time required by the PCA baseline. This validates ALA as the superior, highly scalable mechanism for real-world execution. Visual illustration for the results in Table 3.1 for HAR with ALA are listed in Appendix B.2.2.

6- Differential Privacy (DP-SGD) In our HAR dataset evaluation, Differential Privacy (DP) with medium noise budget setup $(32, 10^{-5})$ drastically decreases model testing accuracy by 34.7% while yielding only a marginal 2.7% decrease in average Source Inference Attack (SIA) accuracy. Furthermore, DP worsens the FI(SIA$_{acc}$) by 11.1% compared to the baseline, demonstrating that despite slightly increasing overall protection, it fails to improve empirical privacy fairness. Highly skewed, non-IID datasets make equitable privacy pro-



(a) Server-side aggregation weights ($\mathbf{w}$). (b) Client-side relative overfitting rank (ρ_k).

Figure 3.6: The dynamic interplay of F*in*P's closed-loop architecture over communication rounds in HAR.

tection fundamentally challenging. Specifically, DP induces significantly higher client losses compared to our F*in*P method, driving the drastic utility drop. Ultimately, even though DP forces client losses to become uniformly high and closely clustered, empirical privacy fairness does not improve. Visual illustration for the results in Table 3.1 for HAR with DP-SGD are listed in Appendix B.2.3.

Finally, empirical analysis of the local loss landscapes (Figure B.13 in Appendix B.2.4) validated our theoretical foundation. We observed a near-perfect Spearman's rank correlation (≈ 1) between the top Hessian eigenvalue ($\lambda_{\max}$) and the Hessian trace (H_T). Because both metrics reliably indicate local overfitting, they are equally weighted in our calculation of the relative overfitting rank (Equation 3.2).

3.2.5.4 Evaluating F*in*P on the CIFAR-10 Dataset

While the HAR dataset demonstrates F*in*P's efficacy on human-centric sensor data, we utilize the CIFAR-10 dataset (distributed via severe Dirichlet sampling, $\alpha = 0.5$) to **stress-test our framework under mathematically controlled, extreme non-IID visual settings.** More on the dataset setup is in Appendix B.1. We evaluate F*in*P against both standard FedAvg and FedAlign Mendieta et al. [2022], a state-of-the-art FL methodology specifically designed to address data heterogeneity. The core results using the ResNet architecture are

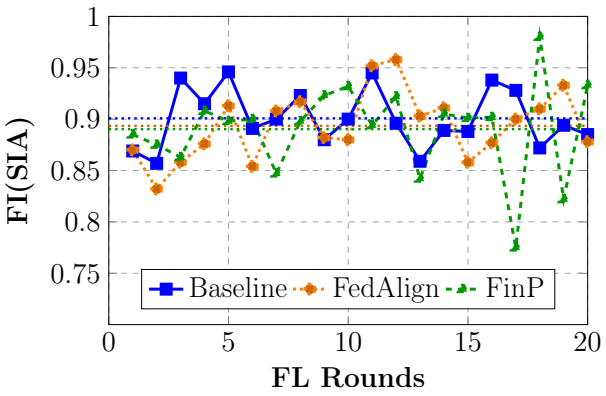


Figure 3.7: Disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with ResNet and F*in*P with PCA.

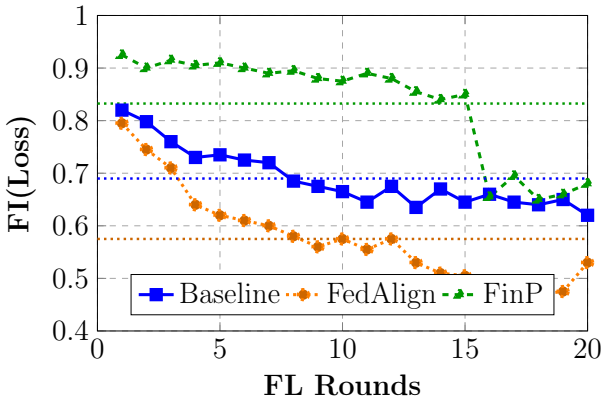


Figure 3.8: Disparity of prediction loss FI(Loss) among clients using CIFAR-10 dataset with ResNet and F*in*P with PCA.

summarized in Table B.1 in Appendix B.3 and explained below.

1. Outperforming State-of-the-Art Heterogeneity Defenses A critical finding of our evaluation is that standard methods for handling non-IID data do not inherently resolve privacy disparities. Despite employing sophisticated local distillation techniques, the FedAlign baseline failed to effectively mitigate SIA risks, exhibiting a higher vulnerability variance (CoV(Loss) = 0.86) than even standard FedAvg (0.67).

In contrast, F*in*P directly targets the loss landscape, achieving a Fairness Index for target prediction loss (FI(Loss)) of 0.83. This represents a substantial 20.3% improvement over FedAvg (0.69) and a 43.1% improvement over FedAlign (0.58). By structurally flattening the inter-client loss differences, F*in*P successfully neutralizes the high-confidence statistical signals emitted by outlier models (further illustrated in Figures 3.7 and 3.8).

2. Approaching the Theoretical Minimum for Absolute Privacy Beyond enforcing fairness, F*in*P dramatically reduces the absolute privacy leakage across the federation. For a balanced 10-class classification task like CIFAR-10, the theoretical random-guess probability for a source inference attack is 1/10 (10%).

As shown in Table B.1 in Appendix B.3, while FedAvg and FedAlign suffer from Mean(SIA_{acc})

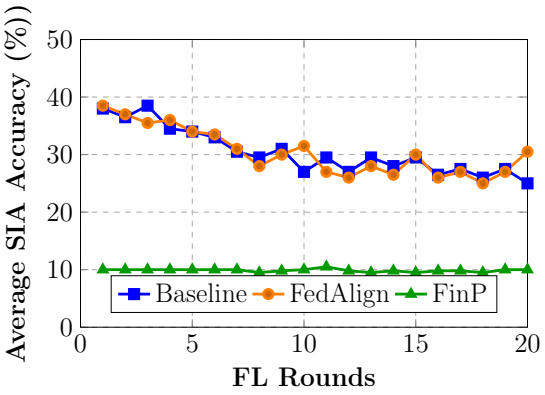


Figure 3.9: Average SIA accuracy in CIFAR-10 with ResNet with F*in*P with PCA.

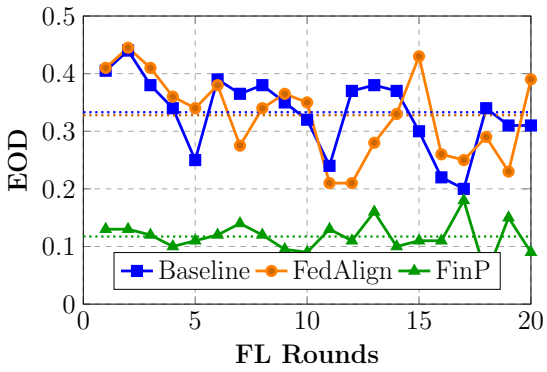
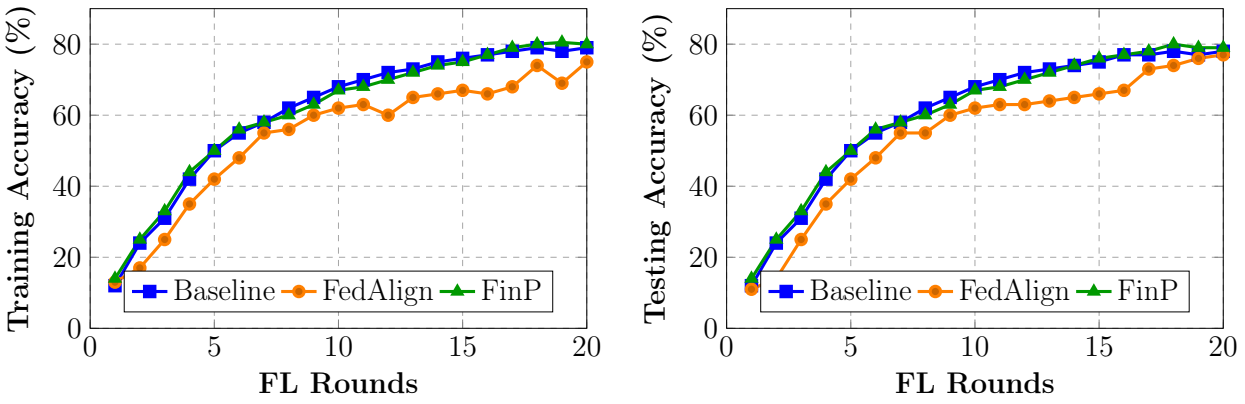


Figure 3.10: EOD in CIFAR-10 with ResNet with F*in*P with PCA.

rates of approximately 30.86%, F*in*P successfully drives the Mean(SIA_{acc}) down to 10.07%—effectively reducing the adversary's attack success to a random guess. Furthermore, the maximum vulnerability observed for any single client (Max(SIA_{acc})) is reduced from 38.52% to 10.67%.

3. Enhancing Group Privacy Fairness (Equal Opportunity) F*in*P maintains comparable baseline CoV(SIA_{acc}) and FI(SIA_{acc}) metrics but demonstrates a massive improvement in Equal Opportunity Difference (EOD). F*in*P reduces the EOD from 0.28 (in both baselines) down to 0.12. As visualized in Figures 3.9 and 3.10, this 57.14% reduction in EOD mathematically guarantees that the adversary's True Positive Rate is vastly more uniform, preventing the systematic exploitation of distinct visual outliers.

4. The Synergy of Privacy and Utility Traditionally, strong privacy defenses degrade model utility. However, F*in*P achieves a rare synergy: it actually *improves* global classification accuracy. F*in*P achieves a testing accuracy of 78.46%, marginally higher than FedAvg's 77.62% (requiring only two additional FL communication rounds to converge, as seen in Figure 3.11). This 0.84% improvement proves our core hypothesis: by utilizing Lipschitz regularization to actively penalize localized memorization, F*in*P forces the constituent local models to learn generalized features, thereby simultaneously improving both privacy and global model utility.



(a) Training accuracy.

(b) Testing accuracy.

Figure 3.11: Global model classification accuracy using CIFAR-10 with ResNet with F*in*P with PCA.

5. Ablation on Client-Side Impact Factor (β) To analyze the sensitivity of the dynamic regularization term, we conducted an ablation study on the impact factor $\beta \in \{0.05, 0.1, 0.3, 0.5\}$ using a standard CNN architecture (Table B.2 in Appendix B.3). The parameter β controls the baseline strength of the Lipschitz penalty applied to the local training loss.

- **Optimal Generalization ($\beta = 0.3$):** Compared to the baseline, increasing β to 0.3 optimally balanced fairness and utility. It improved FI(Loss) from 0.83 to 0.87 and reduced absolute Mean(SIA_{acc}) from 40.91% to 31.85%, while simultaneously improving testing accuracy due to enhanced generalization.

- **Over-regularization Failure ($\beta = 0.5$):** We observed that excessive penalization dominates the client's local training loss, preventing the models from learning valid feature representations. At $\beta = 0.5$, the global model completely failed to converge (Figure 3.15). While this failure state technically may yield the lowest average SIA accuracy (Appendix B.3 Figure B.14), this represents a catastrophic loss of utility rather than a legitimate privacy defense, underscoring the necessity of tuning β to an appropriate task-specific threshold.

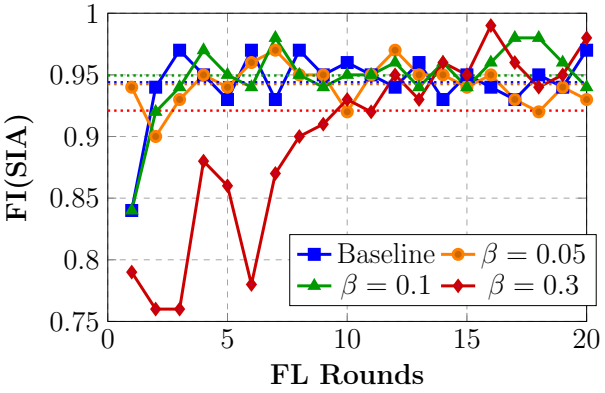


Figure 3.12: Disparity of SIA accuracy FI(SIA) with PCA among clients using CIFAR-10 dataset with base model CNN and F*in*P with PCA.

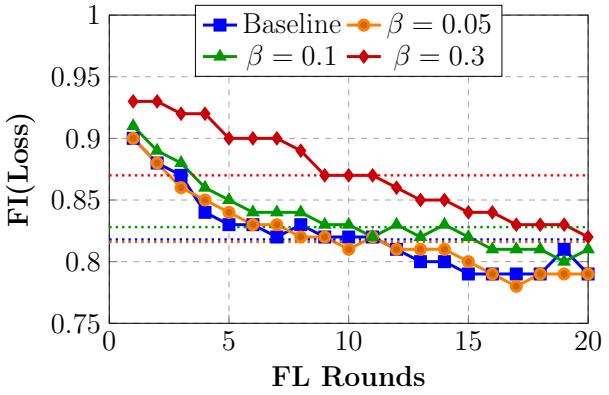


Figure 3.13: Disparity of prediction loss (FI(Loss)) with PCA among clients using CIFAR-10 dataset with base model CNN and F*in*P with PCA.

6. Scalability: Adaptive Lightweight Aggregation (ALA) vs. PCA Finally, to validate the deployment practicality of F*in*P, we substituted the computationally prohibitive PCA-based server aggregation with our proposed Adaptive Lightweight Aggregation (ALA) mechanism (Table B.2 in Appendix B.3).

As formalized in Section 3.2.4.2, ALA completely bypasses expensive high-dimensional PCA distance calculations. Instead, the server directly utilizes the relative overfitting rank (ρ_k)—which it already computes from the incoming model weights—as a direct inverse-weighting scalar for aggregation.

Our results confirm that ALA maintains seamless model convergence within 20 rounds (Figure 3.16). Crucially, ALA achieves comparable improvements in FI(Loss) (Figure 3.18) and SIA reduction (Figure 3.19) as the heavy PCA baseline. This proves that ALA provides identical structural privacy fairness without introducing any computational bottlenecks at the server, rendering F*in*P highly practical for real-world, large-scale FL ecosystems.

7. Differential Privacy (DP-SGD) Our evaluation on the CIFAR dataset yields results consistent with those observed on HAR: DP affords minimal empirical protection against SIA while incurring a drastic degradation in testing accuracy and further exacerbating privacy unfairness. A more detailed analysis of this dynamic is provided in Discussion Section 3.2.6.

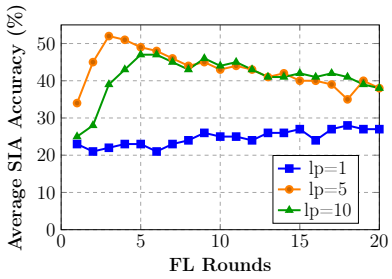


Figure 3.14: Average SIA accuracy on CIFAR-10 (CNN) across varying local training epochs. Increased local memorization directly correlates with higher vulnerability, validating our focus on overfitting.

Visual illustration for the results in Appendix B.3 Table B.2 for CIFAR-10 with DP-SGD are listed in Appendix B.3.2.

8. Validating the Root Cause: Overfitting and Local Epochs Our core thesis posits that localized overfitting—driven by data heterogeneity—is the fundamental mechanism enabling SIAs. Prior literature suggests that increasing the number of local training epochs deepens this memorization, thereby exacerbating privacy risks [Hu et al., 2023, Yeom et al., 2018]. To mathematically validate this within our threat model, we conducted ablation experiments varying the local training epochs ($lp \in \{1, 5, 10\}$) on the CIFAR-10 dataset using a standard CNN. As illustrated in Figure 3.14, the vulnerability scales linearly with local memorization. The Mean(SIA_{acc}) and Max(SIA_{acc}) surged from 24.31% and 27.20% (at $lp = 1$) to catastrophic levels of 42.9% and 51.4% (at $lp = 10$), respectively. This definitively proves that deeper local convergence inherently sharpens the loss landscape, generating high-confidence signals for the adversary. By attacking this exact symptom via client-side Lipschitz regularization, F*in*P successfully neutralizes the root cause of the privacy disparity.

3.2.5.5 Evaluating F*in*P on FEMNIST Dataset

To further validate our findings, we extend our evaluation of the FinP framework to the FEMNIST dataset. Consistent with the results observed on previous datasets, FinP demonstrates aligned performance in this setting. To simulate a strictly non-IID federated environment, we partition the dataset such that each client is exclusively assigned the data of a single,

Table 3.2: Results of FEMNIST dataset with SIA attack.

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics					Efficiency
	Train	Test	Mean (SIA_{acc})↓	Max (SIA_{acc})↓	CoV(SIA_{acc}) /FI(SIA_{acc})	CoV(loss) /FI(Loss)	EOD↓	MAD (loss)↓	Welfare (SIA)↑	Conv. round
Baseline Hu et al. [2023]	72.71	65.04	39.26	49.70	0.305/0.911	0.281/0.916	0.40	0.636	0.546	10
F*in*P ($\beta = 1$)	60.17	60.11	**30.69**	**39.06**	0.408/0.855	**0.208/0.951**	0.41	0.589	**0.626**	12
F*in*P ($\beta = 0.75$)	67.63	63.76	35.77	46.49	0.319/0.905	0.241/0.937	**0.37**	0.597	0.581	12
F*in*P ($\beta = 0.5$)	**67.81**	**64.23**	37.02	45.58	**0.305/0.912**	0.242/0.936	0.37	**0.583**	0.570	14
DP-SGD (ϵ, δ)										
(151, 10^{-5})	51.42	49.97	29.67	34.44	0.272/0.927	0.276/0.926	0.26	1.968	0.661	14
(92, 10^{-5})	28.15	29.41	21.35	25.90	0.313/0.910	0.234/0.947	0.22	1.656	0.750	-
(31, 10^{-5})	4.66	5.68	12.70	15.66	0.555/0.764	0.398/0.860	0.24	9.062	**0.835**	-

unique writer. Our results demonstrate that FinP successfully mitigates Source Inference Attacks (SIA), decreasing mean SIA by 8.57% and max SIA by 10.64% in Table 3.2, while maintaining comparable model accuracy to the baseline.

To comprehensively evaluate fairness improvements, we introduced 2 additional metrics, *Mean Absolute Difference(MAD)* and *Sen's social welfare Sen [1997].* Empirical results indicate that FinP significantly curtails vulnerability to Source Inference Attacks (SIA), yielding reductions of 8.57% in mean SIA and 10.64% in max SIA, without compromising global model accuracy relative to the baseline. Beyond aggregate performance, our evaluation demonstrates that FinP actively mitigates privacy disparities across the federated cohort. Specifically, the framework reduces the Mean Absolute Difference (MAD) by up to 0.053 and increases Sen's Welfare by 8% as demonstrated from Table 3.2.

In contrast, when compared to standard Differential Privacy (DP) methods, we observe that DP significantly degrades global model accuracy while simultaneously exacerbating the MAD. A larger MAD indicates greater variance in client-level loss, which inherently boosts an attacker's linking prediction confidence for highly vulnerable subsets. Consequently, although DP methods may decrease overall SIA accuracy at the cost of utility, they ultimately worsen the disparity in privacy risks, leaving certain clients disproportionately exposed. F*in*P successfully avoids this tradeoff, maintaining global utility while ensuring a highly equitable distribution of privacy preservation. Detailed definitions of these metrics, further analysis and visual results, along with Hessian computation overhead, are provided in Appendix B.4.

3.2.5.6 Evaluating MIA

We also evaluate client privacy under a White-Box Membership Inference Attack (MIA), assuming an honest-but-curious central server. By injecting a dummy client to act as a memorization baseline, the server dynamically trains a Random Forest classifier at each communication round using the layer-wise cosine similarity between target gradients and mock local weight updates. Our evaluations demonstrate that the FinP framework inherently defends against this attack, reducing MIA accuracy by an average of 5.15% compared to the baseline by fostering a more generalized model training process that mitigates localized overfitting. Full details regarding the data partitioning, shadow training generation, the per-round mia evaluation pipeline and visual results are provided in Appendix B.5.

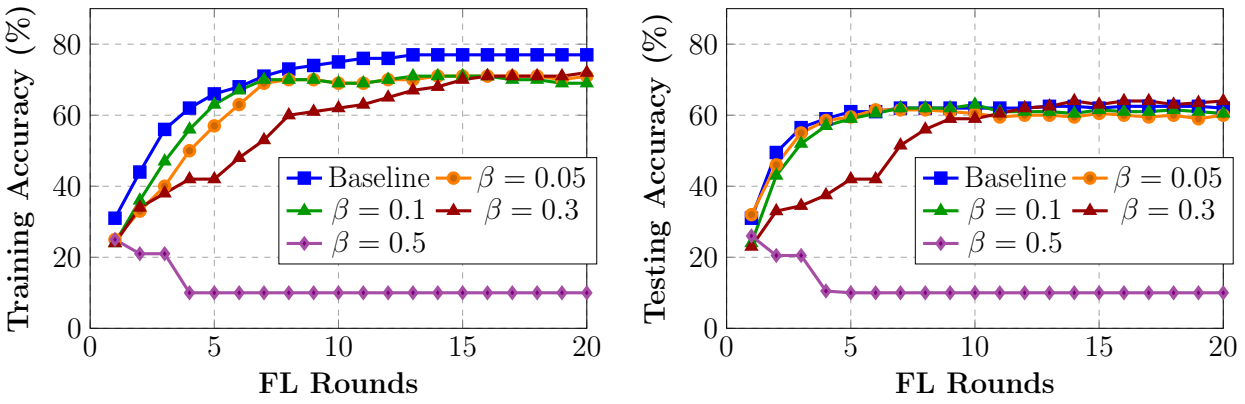
3.2.5.7 Summary of Empirical Results: F*in*P across HAR and CIFAR-10

Across our diverse evaluations—spanning human-centric mobile sensor data (HAR) and mathematically extreme non-IID visual distributions (CIFAR-10)—F*in*P consistently demonstrates its capacity to enforce structural fairness in privacy without sacrificing global model utility.

While the HAR dataset presented a naturally saturated baseline for prediction loss fairness, F*in*P successfully formalized and bounded this equity, proving its stability and robustness in highly sensitive, real-world edge environments without causing detrimental utility trade-offs.

The most definitive evidence of our framework's efficacy emerges in the CIFAR-10 evaluation. Under severe, mathematically controlled data heterogeneity, F*in*P achieves two critical milestones for federated privacy:

- **Neutralizing Absolute Vulnerability:** By actively penalizing localized memorization through dynamic Lipschitz regularization, F*in*P drives the average SIA success rate down to 10.07%, effectively neutralizing the honest-but-curious adversary's capability to a math-



(a) Training accuracy.

(b) Testing accuracy.

Figure 3.15: Ablation experiment of global model classification accuracy with F*in*P with PCA using CIFAR-10 with CNN. The convergence round (conv. round) denotes the point at which the accuracy stabilizes and ceases to make significant improvements in testing accuracy.

ematical random guess.

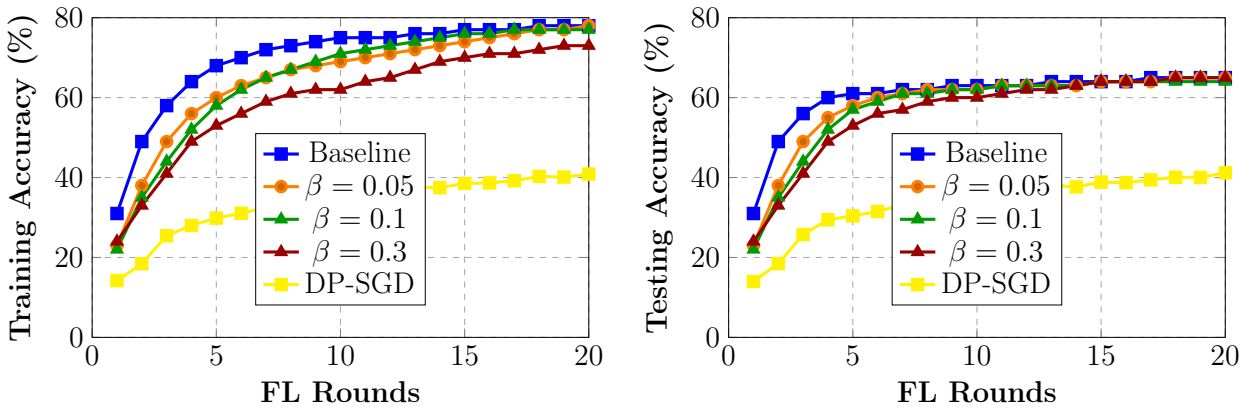
- **Enforcing Equitable Protection:** Simultaneously, the framework achieves a remarkable 57.14% improvement in the Equal Opportunity Difference (EOD). This structurally eliminates the privacy disparities that traditionally plague non-IID federations, ensuring that statistical outliers (minority data distributions) are not disproportionately targeted.

Ultimately, F*in*P shifts the paradigm of federated defense from reactive, post-hoc mitigation to proactive generalization. By flattening the inter-client loss landscape and denying the server the high-confidence signals required for source attribution, our framework guarantees that high model utility and equitable privacy can co-exist in highly heterogeneous networks.

3.2.6 Discussion, Limitations, & Future Work

Analysis on Differential Privacy (DP) against SIA.

F*in*P can be viewed as a complement to Differential Privacy (DP) rather than a replacement: while DP provides formal, worst-case mathematical guarantees against inference attacks, F*in*P specifically targets the fairness of privacy distribution across clients — a dimension



(a) Training accuracy.

(b) Testing accuracy.

Figure 3.16: Adaptive lightweight aggregation (ALA) of global model classification accuracy using CIFAR-10 with CNN.

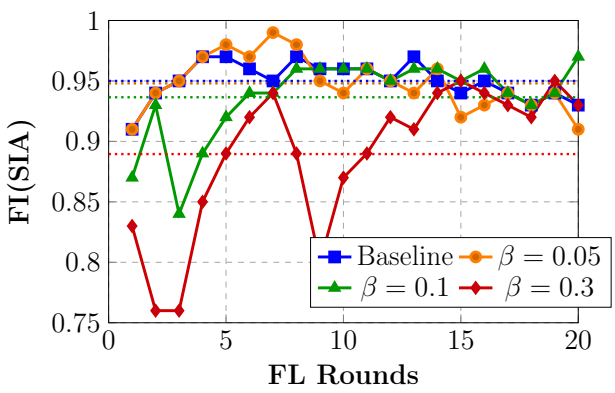


Figure 3.17: Adaptive lightweight aggregation (ALA) disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with base model CNN.

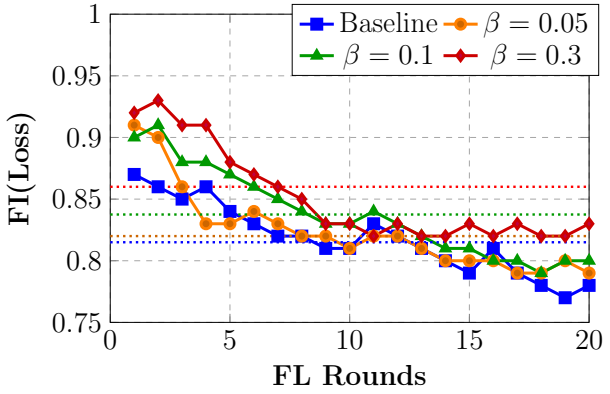


Figure 3.18: Adaptive lightweight aggregation (ALA) disparity of prediction loss FI(loss) among clients using CIFAR-10 dataset with base model CNN.

that DP does not explicitly address. Each approach is preferable under distinct conditions. DP is the appropriate choice when formal, auditable privacy guarantees are required regardless of utility cost, such as in regulated environments subject to GDPR compliance fin [2018]. F*in*P, by contrast, is preferable when the primary concern is equitable risk distribution in highly non-IID, resource-constrained settings, where DP's geometry-blind uniform noise injection creates severe utility degradation without resolving per-client privacy disparities. As demonstrated empirically (Table 3.1), DP-SGD degrades global model testing accuracy by 34.7% while yielding only a marginal 2.7% decrease in average SIA accuracy and actually worsening $FI(SIA_{acc})$ by 11.1% compared to the baseline — inadvertently shield-

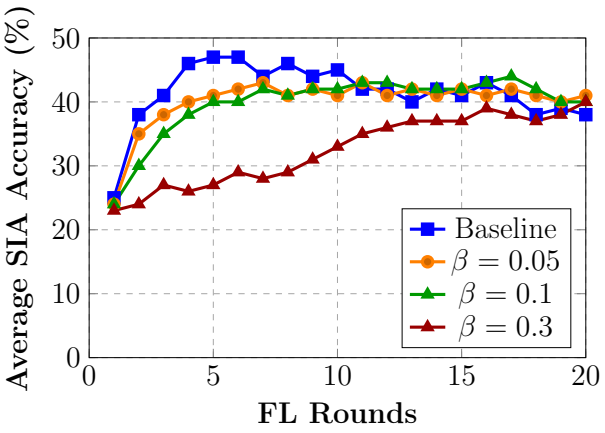


Figure 3.19: Adaptive lightweight aggregation average SIA accuracy of Baseline and different β using CIFAR-10 dataset with CNN.

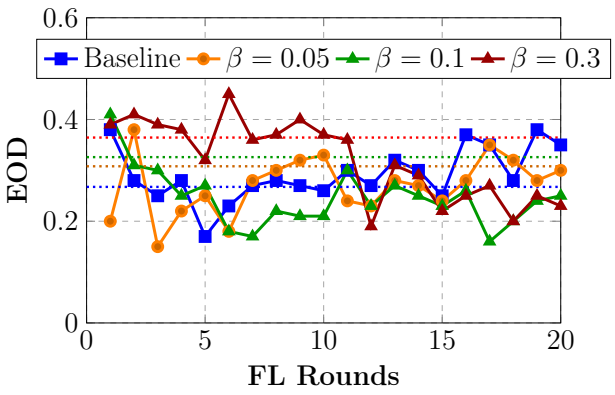


Figure 3.20: Adaptive lightweight aggregation equal opportunity difference (EOD) using CIFAR-10 dataset with CNN.

ing consensus-aligned clients while failing to obscure the highly directional updates of the most disparate clients. Although FinP does not provide formal worst-case privacy guarantees equivalent to DP, it successfully reduces the Mean Absolute Difference (MAD) in loss across clients, a metric that DP tends to exacerbate. In highly non-IID settings, DP applies uniform protection such as fixed clipping and noise across heterogeneous clients, which disproportionately degrades performance for outlier clients and increases the overall MAD in loss. This widened loss disparity can inadvertently boost an attacker's confidence in exploits that rely on loss differences across clients. By reducing this MAD, FinP mitigates this specific empirical vulnerability. Results on FEMNIST show consistent findings in Appendix B.4 Table 3.2. Furthermore, adaptive clipping variants like DP-Fair Yang et al. [2023], designed to improve utility fairness, paradoxically exacerbate the SIA threat by preserving the exact geometric fingerprints the attack exploits. Recent work by Bendoukha et al. Bendoukha et al. [2025] further confirms this inherent tension between fairness interventions and privacy leakage in federated learning. In contrast, F*in*P resolves this disparity without relying on post-hoc noise by locally enforcing a Lipschitz constraint to suppress memorization at its root and globally applying a curvature-aware aggregation penalty to down-weight the most vulnerable updates. Looking forward, combining DP and F*in*P represents a highly promising direction: by dynamically adjusting per-client noise budgets based on local loss

sharpness, future systems could design an adaptive, geometry-aware DP framework that achieves formal DP bounds and empirical privacy fairness simultaneously while minimizing utility degradation.

Limitations F*in*P assumes honest clients submitting genuine updates and metrics. If a client were actively malicious, it could submit falsified loss-sharpness metrics to manipulate the server's adaptive aggregation weights, or intentionally generate overfitted updates to inflate its assigned vulnerability rank. Defending against such data and metric poisoning would require integrating Byzantine-robust aggregation frameworks, which is beyond the current scope of FinP. We acknowledge this as a practical limitation and plan to explore robustness enhancements in future work. F*in*P targets privacy leakage from localized memorization and overfitting, which is the primary driver of SIA disparities in non-IID federated learning. It does not provide formal worst-case privacy guarantees equivalent to Differential Privacy. For attacks independent of memorization — including Property Inference Attacks targeting macro-level statistical properties and gradient reconstruction attacks exploiting raw gradient updates during transmission — F*in*P's efficacy is not guaranteed, as flattening the loss landscape does not inherently prevent these vectors. These limitations, along with the dependence on honest-client and honest-but-curious server assumptions, define the conditions under which F*in*P is effective and should be considered when deploying the framework in settings with stronger adversarial assumptions.

Future Work and Broader Impact

Our current definition of privacy disparity is evaluated in standard, singular-decision FL setups. Future work will explore mapping F*in*P to other gradient-inversion attacks and sequential decision-making environments (e.g., reinforcement learning), where privacy risk must be measured over a continuous trajectory of interactions rather than a static dataset Zhao et al. [2024]. Finally, we aim to explore the integration of our framework with orthogonal Privacy-Enhancing Technologies, such as Secure Multi-Party Computation (SMPC) or

lightweight cryptographic masking, to provide defense-in-depth guarantees for decentralized AI and inform future data governance frameworks.

3.2.7 Conclusion

The F*in*P framework addresses a critical vulnerability in modern Federated Learning: the inequitable distribution of privacy risks. While traditional FL implementations are optimized to preserve average, network-wide privacy, they systematically ignore the severe disparities inflicted upon statistical outliers. Driven by non-IID data distributions, these distinct users suffer from extreme localized overfitting, rendering their updates highly recognizable and forcing them to bear a disproportionate burden of the network's privacy leakage.

This dynamic raises profound ethical concerns in collaborative learning ecosystems, as it directly translates data heterogeneity into structural discrimination against minority users. F*in*P successfully resolves this by shifting the paradigm from reactive defense to proactive generalization. Through the synergistic combination of server-side adaptive aggregation and client-side collaborative overfitting reduction, our framework flattens the inter-client loss landscape. By structurally denying the adversary the high-confidence signals required for source attribution, F*in*P reduces absolute attack success while guaranteeing a fundamentally equitable distribution of privacy for all participants.

3.3 Discussion

Privacy in federated learning is a *distribution* problem as well as a *magnitude* problem. Most privacy machinery asks how much information leaks and tries to make that number small on average or in the worst case. FinP asks *to whom* the leakage accrues and treats the spread of risk across people as the quantity to be governed. This is the same shift Chapter 2 applied to experience, now applied in a domain where equity arguments are rarely made.

Two consequences follow from the approach, independently of any single experiment. Distributing risk fairly does not conflict with protecting privacy in absolute terms: constraining the clients that memorize most reduces overall risk as well as disparity, so reducing disparity also lowers the overall success of the attack. A mechanism that ignores per-client geometry cannot deliver this. Uniform differential privacy adds the same noise everywhere, so it cannot selectively obscure the distinctive updates of outlier clients without paying a heavy, indiscriminate utility cost; the equity gain here comes from being geometry- and client-aware. FinP and formal privacy mechanisms are complementary. FinP provides empirical equity; formal mechanisms provide mathematical guarantees. They can be combined through geometry-aware, per-client budgets.

Chapters 2 and 3 have both pursued fairness *across a group of people*, first in the decisions a system makes and then in the privacy risk it distributes. Both treat people as a set. Even so, a system can distribute experience and risk evenly and still misunderstand what each individual wants. Chapter 4 turns from equity of treatment across a group to faithfulness of understanding for one person: how can a system adapt to a particular person's preferences from a reliable signal about who they are?

Chapter 4

Personalization through Psychometric Traits

4.1 Overview

Personalization asks how well a system can adapt to *one* person, as distinct from how outcomes are distributed across many. This chapter also moves from cyber-physical systems to the conversational large language models (LLMs) that increasingly mediate individual recommendations and decisions.

The motivating difficulty is an asymmetry between how LLMs and humans handle accumulated interaction. As a conversation grows, an LLM's ability to follow a user's stated preferences degrades; prior work reports preference-following accuracy collapsing after only a handful of turns and sliding further as history accumulates. Adding more memory, a larger context window, or a fuller transcript makes the problem worse. Humans work in the opposite direction: the more we interact with someone, the *more* reliable our model of them becomes, because we do not replay every sentence they have ever said. We abstract, forming

a compact sense of who a person is and generalizing from it. This chapter addresses that gap.

The proposal is to supply the abstraction that LLMs lack, and to ground it in psychology, not in surface topic. The chapter's paper, **PACIFIC** [Zhao et al., 2026a], uses the Big-Five (OCEAN) personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) as a stable, compact latent signal for organizing and reasoning over a user's history. Personality connects preferences *across* domains through shared psychometric alignment, not through surface-level semantic similarity: a consistently introverted user's choices in music, travel, and social plans cohere because they share a trait, not because they share keywords. Retrieval that is topically right can therefore be psychologically wrong, latching onto the word "mixer" in a history and recommending a party to someone whose history also shows they leave gatherings early. PACIFIC investigates three questions: whether Big-Five traits are a useful abstraction that lets LLMs infer preferences on a new query; how personality information should be delivered to maximize its benefit; and, because real deployments have no ready trait labels, whether a retriever can surface personality-aligned context without them. The third question is answered by **PiRAG**, a persona-aware retriever that fine-tunes a standard dense retriever with a contrastive objective, pulling together preferences from the same trait and pushing apart mismatched ones, so that retrieval favors persona-consistency over topical proximity. The work is supported by a purpose-built, balanced psychometrics dataset spanning the full high/low spectrum of all five traits, annotated with a dual strategy that mirrors trait level for four of the traits but treats neuroticism on a safety-versus-efficiency criterion, since a high-neuroticism user seeks reassurance rather than more of the stressor.

Section 4.2 introduces PACIFIC, followed by the discussion in Section 4.3.

4.2 PACIFIC: Can LLMs Discern the Psychometric Traits Influencing Your Preferences? Personality-Driven Preference Alignment in LLMs

4.2.1 Introduction

As Large Language Models (LLMs) become increasingly integrated into everyday applications, users naturally expect them to infer their preferences when facing new questions, even when the relevant preference has never been explicitly discussed. At the same time, continued interaction produces rapidly growing user history, making long-term personalization more difficult. Although modern systems such as Gemini [Team et al., 2023] and GPT-4 [Achiam et al., 2023] have substantially improved language understanding and generation, they still struggle to identify and use the right preference signals over long conversations. Prior work [Zhao et al., 2025] finds that preference-following accuracy can fall below 10% after only 10 turns (about $3k$ tokens) for many models, and performance continues to decline as the conversation grows.

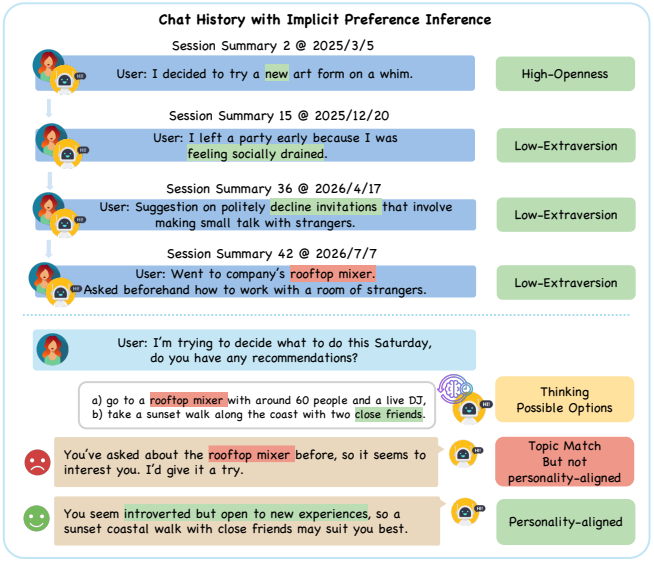


Figure 4.1: Motivation: Topic matching may fail to capture a user's underlying preferences, whereas personality-based reasoning enables better preference inference in a new query. The illustration above shows a simplified example.

Interestingly, humans often process long interaction history in the opposite way: as interactions accumulate, their understanding of another person can become more reliable, abstract, and generalizable. Consider the example in Fig. 4.1: a friend you know tends to *try new*

experiences, feel socially drained at parties, and avoid small talk with strangers. If they later ask for suggestions for a Saturday activity, a topic-based strategy may focus on their earlier mention of a *rooftop mixer* and recommend attending it. As a human, however, you may instead easily infer from the interactions that they would likely prefer trying something new alone or with a few close friends. These seemingly unrelated past behaviors reveal a consistent tendency toward introversion. In other words, people do not recall previous conversation details; instead, they form a higher-level understanding of another person's personality and use it to generalize to new situations.

This suggests that effective LLM personalization may require more than storing and retrieving preference statement histories. Instead, LLMs should learn to use these histories to construct a more structured representation of the user and use that representation to infer user preferences and give recommendations for questions from new domains.

We explore *Big-Five personality* as such a representation. The Big-Five personality traits – Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) – capture relatively stable behavioral tendencies and are associated with systematic patterns in preferences and decision making Li et al. [2025]. We propose that, ***rather than requiring an LLM to remember many specific preferences, we use Big-Five traits as a more stable signal that helps organize and interpret preferences across domains.*** This allows preferences to be connected through shared psychometric alignment rather than semantic similarity alone. We therefore ask: *Can personality serve as a useful latent representation for organizing user preferences and improving preference following in LLMs?*

We first test our hypothesis on PrefEval Zhao et al. [2025], an independently constructed preference-following benchmark of 1,000 synthesized user preferences, questions, and ground-truth answers. Our results (Tab.4.1) show that personality-aligned preference information improves preference following of LLMs.

However, PrefEval was not designed for systematic personality analysis, and its preferences are unevenly distributed across the ten high/low Big-Five trait directions. To enable controlled comparisons, we construct a balanced psychometrics-based dataset and investigate the following research questions illustrated in Fig. 4.2:

- **RQ1 (Personality-Driven Preference Alignment).** Can Big Five traits provide a useful user abstraction that enables LLMs to infer preferences in a new query especially when that preference has never been explicitly stated?

- **RQ2 (Effective Use of Personality Information).** How should personality information be incorporated to maximize its benefit for preference prediction?

- **RQ3 (Retrieving Personality-Aligned Information).** Without explicit personality labels, can a retriever identify personality-aligned memories that support the prediction of unseen preferences?

We make three contributions: (1) We show that personality traits improve preference following in personalized question answering. (2) We propose the **PACIFIC** framework, which incorporates personality-aligned preferences during generation. LLMs reach near-ceiling accuracy given clean trait-aligned context, and since real histories are messy and unlabeled, our persona-aware retriever (PiRAG) improves label-free accuracy from 30% to 43% without trait annotations. (3) We introduce a psychometrics-based dataset with 1,200 preference statements and conversations spanning diverse domains (e.g., travel, movies, education), annotated with Big-Five (OCEAN) trait directions [Briggs, 1992, De Raad, 2000, Goldberg, 2013].

4.2.2 Persona Trait Boosts Preference Following

We evaluate our hypothesis using PrefEval [Zhao et al., 2025], an existing benchmark for preference following. As illustrated by the blue steps in Fig. 4.3, PrefEval covers 20

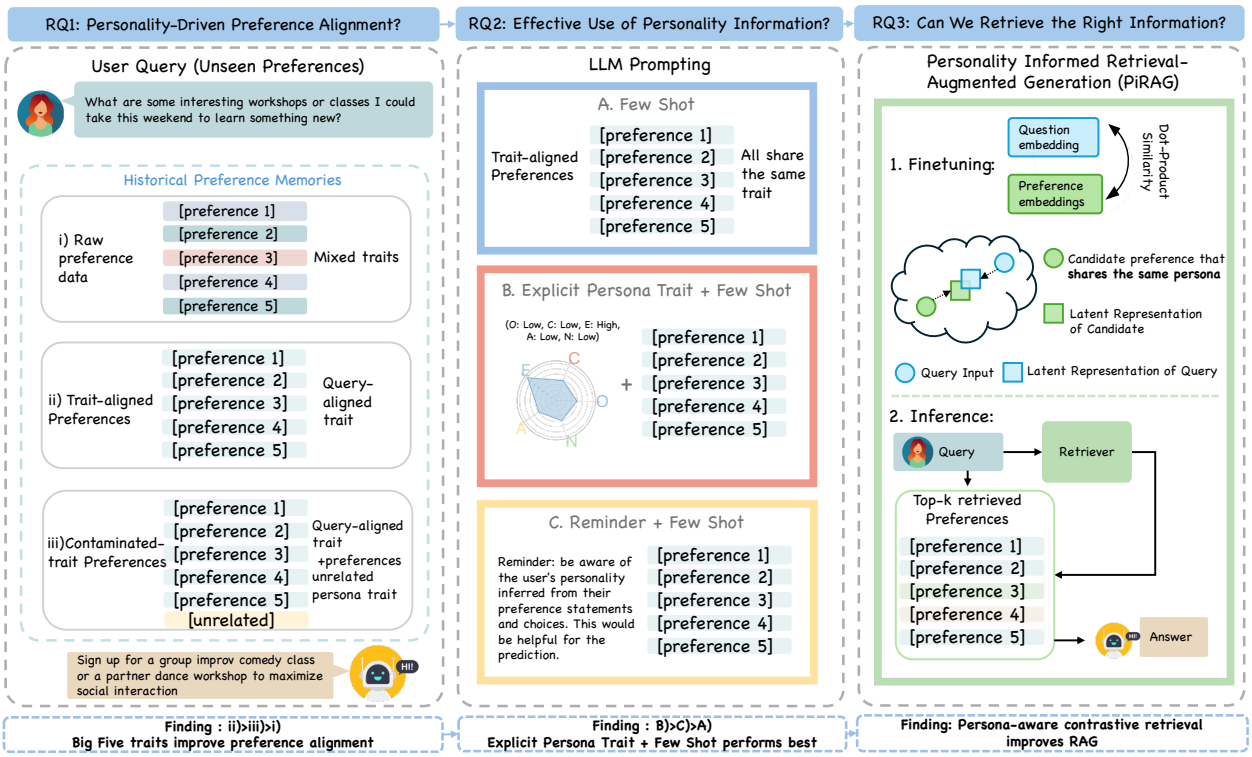


Figure 4.2: **Overview of our PACIFIC framework. RQ1 (left)** evaluates whether personality provides a useful latent structure for preference alignment by comparing mixed-trait, trait-aligned, and contaminated trait-aligned preference histories for predicting an unseen user preference. **RQ2 (middle)** investigates how personality information should be incorporated into the LLM prompting, comparing few-shot alone, with explicit persona traits, and with implicit persona reminders. **RQ3 (right)** studies whether personality-aligned information can be automatically retrieved using Personality-informed RAG (PiRAG).

topics (listed in Appendix C.1.6) and consists of synthesized user preferences p_i, downstream multiple-choice questions q_i, and four candidate answers $\{c_{i,k}\}_{k=1}^{4}$. For each question, $a_i \in [1, 4]$ denotes the index of the preference-adhering ground-truth answer, i.e., c_{i,a_i}. An LLM-based judge is then used to filter out invalid or insufficiently challenging instances, resulting in a final set of 1,000 preference–question–answer triples (p_i, q_i, a_i).

4.2.2.1 Annotate Trait Label

Importantly, we leave these original preferences, questions, and answer labels unchanged and augment each preference statement and each question choice with Big-Five personality annotations, as shown in Step 4 of Fig. 4.3. Specifically, each preference p_i and answer choice

$c_{i,k}$ is scored on a 7-point Likert scale along the five personality dimensions $\{O, C, E, A, N\}$, where 1 indicates a very low expression of a trait and 7 indicates a very high expression.

Dual-Strategy Annotation To ensure psychometric validity, we employ a Dual-Strategy Annotation Protocol separating identity-expressive traits $\{O, C, E, A\}$ from need-compensatory traits $\{N\}$. Let x denote an annotated context, i.e., either a preference p_i or an answer choice $c_{i,k}$. Each such x is annotated with (i) a trait score vector $\mathbf{s}(x) \in [1, 7]^5$ over $\{O, C, E, A, N\}$ and (ii) a confidence vector $\boldsymbol{\gamma}(x) \in [0, 1]^5$ indicating the trait score reliability.

The Mirroring Strategy (O, C, E, A) For Openness, Conscientiousness, Extraversion, and Agreeableness, we adopt a Mirroring Strategy from Self-Congruity Theory [Sirgy, 1985, 2018], which posits that individuals prefer products, activities, or content that reinforce their established self-concept. Users with these traits predominantly seek ***Alignment***; for example, a highly Conscientious user prefers highly structured tools, while a highly Open user prefers novel and complex experiences. Scoring Criteria for O, C, E, A are in Supp. C.1.2.

The Compensatory Strategy (N) For Neuroticism (N), the Mirroring Strategy is psychometrically flawed. A user with High Neuroticism (anxiety-prone) does not seek "High Neuroticism" experiences (risk/stress); rather, they seek ***Safety*** [Tamir, 2005, Hirsh et al., 2012]. Therefore, we apply a Compensatory Strategy based on the Safety vs. Efficiency trade-off. High-N users view preferences as a defensive mechanism, prioritizing *Psychological Cushioning* (guarantees, predictability) to mitigate anxiety. Conversely, Low-N users possess high emotional resilience and capacity for stress. They are outcome-oriented and prioritize efficiency over comfort, often perceiving unrequested safety buffers as *coddling*. For example, consider a traveler choosing between a 45-minute and a 3-hour layover. A Low-N user would prefer the tight 45-minute layover to save time, accepting the risk of rushing as a calculated trade-off, whereas a High-N user would prioritize the 3-hour layover for the peace of mind it provides. Scoring criteria for N are provided in Appendix C.1.2.

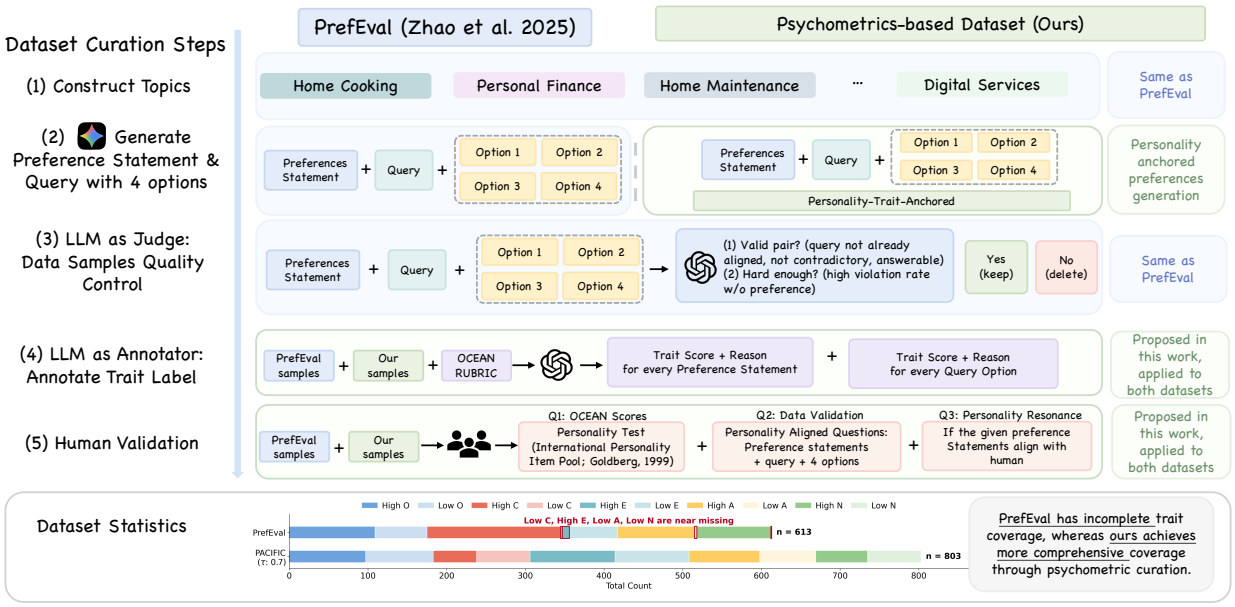


Figure 4.3: **Our psychometrics-based dataset curation pipeline.** We compare the PrefEval Zhao et al. [2025] pipeline with our psychometrics-based extension. Our five-step construction process begins by (1) retaining the 20-topic foundation from PrefEval. (2) To ensure comprehensive trait coverage for targeted analysis, we generate personality-anchored preference statements and queries rather than deriving them directly from topics. (3) Consistent with PrefEval, we filter out statements that language models fail to predict accurately to maintain data quality. (4) We then annotate all statements and query options with OCEAN trait labels and corresponding reasons, and (5) validate them through human assessments of personality, preference alignment, and personality resonance. As illustrated in the bar plot, the resulting dataset achieves significantly more comprehensive coverage across both high- and low-level Big Five traits compared to the PrefEval baseline.

Trait Label Mapping To simplify downstream analysis and reporting, we discretize continuous scores by mapping each context x (e.g., a preference p_i or a choice $c_{i,k}$) to a trait label. Let $\mathcal{D} = \{T^{\mathrm{H}}, T^{\mathrm{L}} \mid T \in \{O, C, E, A, N\}\}$ denote the set of high and low Big Five trait indicators. For each trait, the continuous score $s_t(x) \in [1, 7]$ is mapped to a discrete intensity $d_t(x)$, which is low if $s_t(x) < 4$, high if $s_t(x) > 4$, and unclear if $s_t(x) = 4$. We then assign a categorical label $t(x) \in \mathcal{D} \cup \{\text{unclear}\}$ according to $d_t(x)$: a distinctly high or low intensity yields the corresponding indicator T^{H} or T^{L}, and otherwise $t(x) = \text{unclear}$.

4.2.2.2 Cross-Domain Preference Alignment

We then use these annotations to test whether personality can support cross-domain preference alignment. Specifically, we consider a setting in which the LLM is not given a preference

that directly answers the target question. Instead, it must rely on preferences expressed in other contexts and infer answers from them. For each target question, we construct four preference-context settings (illustrated in Fig. 4.2): *(i) Zero-shot (no preference):* the LLM answers the question without any preference context; *(ii) Mixed Traits:* the LLM receives question-irrelevant preferences drawn from different personality traits; *(iii) Aligned Traits:* the LLM receives question-irrelevant preferences from other domains that share the same underlying personality trait as the target question; and *(iv) Contaminated Aligned:* the LLM receives question-irrelevant but trait-aligned preferences together with additional unrelated or mismatched preferences.

In each setting, we sample 200 multiple-choice questions (20 for each of 10 high/low Big-Five directions). We evaluate preference following using accuracy. For each question q_i, the model selects one of four choices, producing $\hat{a}_i \in \{1, 2, 3, 4\}$, and we compute $\text{Acc}(\%) = \frac{1}{200} \sum_{i=1}^{200} \mathbb{1}\{\hat{a}_i = a_i\} \cdot 100\%$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

We observe two main findings from the results in Tab. 4.1:

1. **Preference context improves personalization. Aligned trait preferences outperform unrelated preferences.** Preference context improves personalization when trait-aligned. In Tab. 4.1, mixed-trait preference history yields only slight improvement over zero-shot (67.50% vs. 46.25%), showing limited benefit when preferences are not aligned with the user's underlying traits. In contrast, trait-aligned preferences substantially improve accuracy from 46.25% to 82.5%, showing that organizing preference history by shared personality traits provides more informative signals for predicting the user's unseen preferences.
2. **Unrelated preferences introduce noise and reduce accuracy.** Comparing (iii) and (iv), adding preferences from other traits lowers accuracy relative to using only trait-aligned preferences (82.5% vs. 72.92%). This indicates that irrelevant preference statements can distract the model and weaken preference-following performance.

However, PrefEval is not evenly distributed across personality traits. As shown at the bottom of Fig. 4.3, its examples provide highly uneven coverage across the ten high/low Big-Five directions, with several traits substantially underrepresented. This makes it difficult to determine whether observed differences across traits reflect genuine personality effects or simply differences in data coverage, preventing reliable trait-level comparisons.

Table 4.1: Overall accuracy (%) on PrefEval across preference-context setups. Mixed and Aligned use five preferences per question (Appendix C.3.1); Contaminated Aligned adds two mismatched preferences as noise (Appendix C.3.2).

	(i) Zero-shot	(ii) Mixed	(iii) Aligned	(iv) Contam.Aligned
Acc. (%)	46.25	67.50	**82.50**	72.92

4.2.3 Psychometrics-based Dataset

To enable systematic evaluation across the full personality spectrum, we construct a psychometrics-based preference dataset covering both high and low expressions of all five Big-Five traits. Unlike the sparse personality coverage of existing PrefEval preference data, our dataset is designed to study how personality relates to preferences and whether the relationship generalizes to unseen decisions.

Grounded in prior research connecting personality with decision-making and preferences [Czerniawska and Szydło, 2021], our pipeline generates 1,200 curated synthetic preference–query pairs spanning 20 diverse domains, including finance, entertainment, travel, and consumer electronics. As illustrated at the bottom of Fig. 4.3, the dataset provides broad coverage across all ten high/low Big-Five directions.

4.2.3.1 Dataset Construction

Generation Pipeline The construction process of our dataset is illustrated by the green boxes in Fig. 4.3. We begin with the same 20 topics used in PrefEval and follow a similar procedure to generate preference–question–answer triples (p_i, q_i, a_i). The key difference is

that our generation process is explicitly anchored to psychometric personality traits. Specifically, we generate examples around each of the ten high/low Big-Five directions, ensuring that the resulting preferences cover 10 personality traits evenly.

To ensure each question genuinely requires personalization, we require at least one incorrect choice $c_{\text{distractor}} \in \{c_{i,k} | k \neq a_i\}$, to satisfy $P(c_{\text{distractor}} \mid q_i) \gg P(c_{\text{distractor}} \mid q_i, p_i)$, where $P(c \mid \cdot)$ denotes how likely c is to be an appropriate answer given the available information. Intuitively, $c_{\text{distractor}}$ appears plausible from the query alone, but becomes clearly inappropriate once the user's preference p_i is known. This ensures that a generic, non-personalized response, conditioned on the query alone, will tend to contradict the user's preference, testing whether a model can proactively apply the latent constraint p_i rather than defaulting to baseline helpfulness.

As in PrefEval, we use an LLM-based judge to filter out invalid or insufficiently challenging samples. The validated preferences may be expressed either explicitly or implicitly through user–LLM interactions. To support end-to-end evaluation in more natural conversational settings, each validated example is further augmented with a brief scenario and a 2–4-turn dialogue that implicitly conveys the target preference. An example is provided in Appendix C.1.1.

Finally, we annotate each preference p_i and each answer choice $c_{i,k}$ with one of the ten personality directions in $\mathcal{D}$, using the same annotation procedure applied to PrefEval, as described in Sec. 4.2.2.1.

4.2.3.2 Dataset Quality Control and Validation

To ensure the high fidelity and difficulty of our psychometrics-based dataset, we implemented a multi-stage validation protocol combining automated filtering, trait-matched human evaluation, and plausibility assessment as shown in Fig. 4.3.

Automated Quality Control To avoid generator-evaluator bias, an independent model (GPT-4o-mini) assessed all instances scalably. Only examples passing two strict rubrics were retained: (1) the ground-truth optimally aligns with the preference while the three distractors violate or ignore it, and (2) the preference logically reflects its assigned trait.

Human Psychometric Grounding To ensure ecological validity, we recruited 15 annotators from diverse professional backgrounds (e.g., PhD students, software and hardware engineers, data analysts) and ran a three-stage study:

1. **Trait profiling.** Annotators completed the Big Five Personality Test [2019], based on Goldberg's Big-Five factor markers [Goldberg, 1992], to record their OCEAN profiles.

2. **Choice validation.** Annotators were assigned 25 randomly sampled preference–question pairs corresponding to their own dominant trait, spanning the ten trait labels in $\mathcal{D}$. Each variant had at least five annotators. When asked to select the generated option best matching the stated preference, they demonstrated strong, intuitive agreement on trait manifestation, achieving an average accuracy of 78.22% and Fleiss' κ of 0.8599 among users with the same assigned questions. Intuitively, this is because annotators were evaluating traits they personally possess. The GPT-4o-mini evaluator reached a Cohen's κ of 0.9170 against this human consensus, validating our automated pipeline.

   This performance of 78.22% is appropriate given that each question includes carefully crafted *hard-negative* distractors to prevent LLMs from exploiting surface-level keywords; the high κ confirms human annotators were highly consistent rather than guessing.

3. **Resonance check.** We assigned annotators preference-question pairs matched to their dominant traits. They then indicated whether they preferred the ground-truth option over the other three options. Choosing the ground truth indicated that this option offered the best alignment for the question. Ultimately, they reported a 91% approval rate.

Distractor Plausibility Assessment To empirically validate our High-Violation constraint, we evaluated an LLM on the dataset queries with the user preference completely withheld. With preference withheld, an LLM scores at random guess chance (25%), confirming all four options are equally attractive absent the personality constraint (Tab. 4.2, zero-shot).

Dataset Statistics For each of the ten high/low Big-Five trait labels, we curate 120 preference–query pairs, in total 1,200 samples. As shown at the bottom of Fig. 4.3, our psychometrics-based dataset provides broader and more balanced coverage across all trait directions than PrefEval, enabling more systematic evaluation across the full personality spectrum.

4.2.4 PACIFIC Framework

We revisit RQ1 (Fig. 4.2) on our psychometrics-based dataset using the same four preference-context settings as before (Zero-shot, Mixed, Aligned, Contaminated Aligned), again **never** providing a preference that directly reveals the answer.

4.2.4.1 Generalizability Across Psychometric Profiles

We evaluate four models: Gemma-3-4B-IT, Llama-3-8B-Instruct, Gemini-2.5-Pro, and GPT-4o-mini. Results for Gemma-3-4B-IT are reported in Tab. 4.2, while results for the remaining models are provided in Appendix C.1.7. We observe the same overall pattern as on PrefEval (Tab. 4.1): trait-aligned preference history substantially outperforms mixed-trait history, while introducing a moderate amount of mismatched preference information causes only a small degradation in performance. This replication on our psychometrics-based dataset provides further evidence that the benefit of personality-aligned preference history is not specific to the particular distribution of PrefEval. Notably, given clean trait-aligned context, strong models approach the ceiling (Gemini-2.5-Pro: 99.25%, Table C.6), confirming

Table 4.2: Overall and trait-wise accuracy (%) across preference-context setups on our psychometrics-based dataset.

	Acc (%)	$\mathbf{O^H}$	$\mathbf{C^H}$	$\mathbf{E^H}$	$\mathbf{A^H}$	$\mathbf{N^H}$	$\mathbf{O^L}$	$\mathbf{C^L}$	$\mathbf{E^L}$	$\mathbf{A^L}$	$\mathbf{N^L}$
(i) Zero-shot	25.75	17.50	25.00	27.50	30.00	37.50	22.50	17.50	35.00	25.00	20.00
(ii) Mixed	29.25	25.00	12.50	2.50	35.00	42.50	40.00	32.50	50.00	37.50	15.00
(iii) Aligned	**63.00**	47.50	42.50	50.00	67.50	50.00	80.00	95.00	85.00	67.50	45.00
(iv) Contam. Aligned	61.75	52.50	35.00	35.00	50.00	60.00	77.50	87.50	87.50	85.00	47.50

the bottleneck is context construction.

More importantly, our psychometrics-based dataset's balanced coverage across personality traits allows us to examine how this effect varies across personality profiles. We evaluate all 32 personality profiles (each trait can take one of two levels (*high* or *low*), for a total of $2^5 = 32$ distinct combination profiles) and measure how much trait-aligned preferences improve performance across personas. We find that personas $(O^L, C^L, E^L, A^L, N^H)$ and $(O^L, C^L, E^L, A^L, N^L)$ achieve the highest accuracy when prompted with trait-aligned preferences (Appendix C.5). Finally, we perform 3-fold cross-validation on our psychometrics-based dataset; the results, reported in Appendix C.6, show the same overall trends as Tab. 4.2.

4.2.4.2 Label-Aware Prompting Strategies

Having established in RQ1 that trait-aligned preference history can substantially improve preference following, we next ask a more practical question: *how should personality information be incorporated into the LLM to best support personalization?*

To isolate the model's ability to reason over persona traits from the complexities of information retrieval, we first assume access to the ground-truth trait labels of both the user's question and their preference statements. As illustrated in RQ2 of Fig. 4.2, we investigate three prompting strategies that expose persona traits at different levels of explicitness, ranging from direct trait labels to implicit reminders:

A. Few-shot (Preference-only) Five trait-aligned preferences are provided as in-context

examples – the aligned setting from RQ1, serving as the baseline (prompt in Appendix C.3.1).

B. Explicit Persona Trait Hints Ground-truth trait labels are exposed in three variants: labeling (1) preferences, (2) preferences and answer choices, or (3) trait labels only with preference text withheld (templates in Appendix C.3.3).

C. Implicit Persona Reminder Strategy A is augmented with a lightweight instruction to consider the user's personality, testing whether encouraging persona-aware reasoning helps without explicit labels (template in Appendix C.3.5).

Together, these strategies compare how effectively trait-label-aware prompting improves LLMs' preference following.

Table 4.3: Accuracy (%) of different ways of adding persona information to preference prompting.

Method	**Variant**	**Acc. (%)**	**Trait-wise Acc. (%)**									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
A. Few-shot	Trait-aligned preferences	63.00	47.50	42.50	50.00	67.50	50.00	80.00	95.00	85.00	67.50	45.00
B. + persona hints	Labeled Prefs	**76.00**	55.00	70.00	72.50	82.50	67.50	87.50	87.50	90.00	90.00	57.50
	Labeled Prefs + choice	57.25	100.00	95.00	100.00	82.50	15.00	40.00	55.00	47.50	17.50	20.00
	Traits only	37.75	100.00	75.00	97.50	67.50	7.50	7.50	2.50	0.00	0.00	20.00
C. + Reminder	Instruction-only (no labels)	67.00	47.50	57.50	50.00	65.00	65.00	82.50	82.50	90.00	75.00	55.00

4.2.4.3 Personality-Informed RAG (PiRAG)

In real-world conversations, personality labels are not explicitly attached to user query statements. A practical personalization system must therefore identify personality-relevant evidence directly from raw interaction history. *Without explicit personality labels, can a retriever identify personality-aligned memories that support the prediction of unseen preferences?*

Label-aware strategies establish that personality alignment helps, but ground-truth trait

annotations are impractical in deployment, where labels are latent. To select trait-consistent context automatically, we propose a retrieval-augmented generation (RAG) framework. We deploy a dual-encoder Dense Passage Retrieval (DPR) architecture [Karpukhin et al., 2020] to map both the user query and the candidate preference statements into a shared dense vector space. Given a query q_i, the system retrieves the top-k preferences based on embedding similarity to dynamically construct the LLM prompt.

Crucially, standard semantic retrieval prioritizes topical relevance over psychometric alignment, without explicitly favoring trait-consistent evidence [Salemi et al., 2024, Wegmann et al., 2022]. To address this, we introduce PiRAG, a persona-aware fine-tuning step for the retriever(shown in RQ3 in Fig. 4.2). We construct contrastive training pairs in which queries and preferences that share the same underlying OCEAN trait form positive pairs, while mismatched traits serve as negatives. This targeted fine-tuning objective pulls the representation space of same-trait representations closer together, encouraging retrieval of personality-consistent user context.

4.2.5 Experiments

4.2.5.1 Setup

Dataset. To evaluate performance across the full personality spectrum, we use our psychometrics-based dataset. We construct disjoint context and test sets as follows:

1. *Context Train Set.* Per trait, randomly draw five preference statements to form the user profile shown in the prompt.

2. *Test Set.* Simultaneously, sample two *non-overlapping* statements to evaluate the model.

3. Repeat steps 1–2 ten times *without replacement*, producing 50 contexts, 20 tests per trait.

4. Run the entire procedure with two random seeds (200 tests/seed) and report the mean.

Prompt format. We use a consistent prompt structure across methods: user preference context followed by the target query. Unless otherwise specified, we include five preference statements in the prompt for each question.

LLM. We evaluate 4 models: two open-weight models, Gemma-3-4B-IT, Llama-3-8B-Instruct, run on the same GPU resources detailed in Appendix C.2; and Gemini-2.5-pro and gpt-4o-mini, run via API. Gemma results are reported in Tab. 4.3, and other models' results can be found in Appendix C.1.7.

Retrieval. We adopt the Dense Passage Retrieval (DPR) architecture [Karpukhin et al., 2020]. We use the standard DPR setup with BERT-based bi-encoders [Devlin et al., 2019]. Fine-tuning details are in Appendix C.4.

4.2.5.2 Results

Overall, accuracy improves significantly when prompts include trait-aligned preferences or implicit persona guidance as shown in Tab. 4.3, demonstrating that organizing preferences via latent traits streamlines personalization. The results of personality alignment via our PACIFIC framework remain consistent across fine-tuning of both open-source small-scale and high-performance, large-scale models (Appendix C.1.7).

Explicit trait labels enhance preference grounding. Adding ground-truth trait labels to preferences gives the highest accuracy (76%; Tab. 4.3 B), improving over few-shot prompting consistently across all ten trait dimensions. This indicates that explicit psychometric cues help bridge stated preferences with downstream choices.

Labeling answer choices triggers positivity bias. Conversely, labeling the candidate answers sharply degrades performance (76% -> 37.75%, 57.25%, Tab. 4.3 B), driven disproportionately by *low*-trait conditions. We attribute this to an LLM positivity bias toward affirmative choices, which penalizes low-trait personas (analyzed in Sec. 4.2.5.3).

Implicit reminders activate latent persona reasoning. A lightweight instruction to consider the user's persona (67%, Tab. 4.3 C) trails explicit labels but clearly beats the preference-only baseline, showing models can activate latent psychometric reasoning without categorical annotations.

Contrastive fine-tuning is necessary for persona-aware retrieval. When trait annotations are absent — the common real-world case — a pretrained DPR retriever yields only 30.25% (Tab. 4.4), since it favors semantic similarity over persona consistency, though this already beats the no-preference and mixed-preference settings (Tab. 4.2). Persona-aware contrastive fine-tuning raises accuracy to 43% (Tab. 4.4; implementation details and trait-wise accuracy results in Appendix C.4), showing the retriever learns to surface trait-aligned preferences. Ground-truth-label prompting remains higher, but requires trait annotations that are impractical at deployment; PiRAG is the label-free counterpart. The 43% is a conservative floor from a deliberately simple DPR retriever; stronger encoders should raise it. Notably, expanding mixed-trait context from 5 to 25 preferences slightly reduces accuracy (from 61.5% to 59.5%; Appendix C.1.7, Tab. C.6), indicating more raw history is not better. We discuss more in Appendix C.7. We further repeat the same experiments on PrefEval and observe consistent trends (Table C.3), further supporting our findings.

Table 4.4: Overall accuracy (%) using a pretrained versus persona-aware fine-tuned retriever.

RAG	Pretrained Retriever	Fine-tuned Retriever(PiRAG)
Acc. (%)	30.25	43.00

Table 4.5: Persona-trait prediction accuracy (%) on 200 samples using Gemma-3-4B-IT.

Input type	**Overall Acc. (%)**	**Trait-wise Acc. (%)**									
		$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
Preferences	54.5	100	100	95	100	100	5	0	0	5	40
Choices	85.69	91.88	63.75	87.5	91.25	75	87.5	85.63	91.25	91.88	91.25

4.2.5.3 Social Desirability Bias

To succeed in realistic, end-to-end conversational settings, models must autonomously infer a user's latent personality from their interaction history. We evaluate this by prompting the LLM to predict OCEAN traits directly from either stated preferences or candidate choices. Tab. 4.5 reveals a counterintuitive discrepancy: trait prediction from abstract answer choices (85.69%) significantly outperforms prediction from explicit user preferences (54.50%).

A trait-wise breakdown exposes the root cause: while the model perfectly identifies "High" traits from preferences, its accuracy collapses (0–40%) on "Low" dimensions. We attribute this to Social Desirability Bias induced by Reinforcement Learning from Human Feedback (RLHF), which can conflate psychometric "low" scores (e.g., a preference for routine over openness) with normatively negative behavior or safety violations[Yi et al., 2025, Salecha et al., 2024]. As a result, models may suppress such signals when judging users. This bias is context-dependent: accuracy recovers when evaluating abstract choices rather than users directly. Consequently, it can bottleneck personalization by favoring socially desirable over preference-aligned recommendations.

4.2.6 Related Work

Personality and Human Preferences. The correlation between the Big Five personality traits and human preferences is well-documented in psychological literature, particularly regarding aesthetic and entertainment preferences [Rentfrow and Gosling, 2003, Rentfrow et al., 2011, Cantador et al., 2013]. Building on these psychological findings, researchers have developed Personality-Aware Recommendation Systems [Hu and Pu, 2011, Ning et al., 2019]. Treating personality as a latent prior for observable preferences motivates our dataset and framework.

Personality, Preference and Large Language Models. Recent work has increasingly focused on the intersection of personality psychology and Large Language Models (LLMs),

shifting from simple instruction following to complex persona simulation. Early research focused on "inducing" specific personality traits into LLMs to make them more human-like [Jiang et al., 2023, Li et al., 2025]. While these works focus on the model's personality, a critical, complementary challenge is extracting the latent knowledge from raw interactions and aligning models with the user's personality. Recent frameworks like Proxona [Choi et al., 2025] address this by distilling audience traits from raw comments into dimensions and values to cluster interactive personas. To effectively serve diverse users, models must understand how latent traits drive behavior, yet no standardized benchmark currently exists to evaluate this reasoning capability. The landscape of existing datasets is primarily dominated by tasks that test explicit adherence rather than psychometric understanding. Prior work [Zhao et al., 2025] focuses on logical constraint satisfaction, and work [Jiang et al., 2025] targets dynamic memory retrieval; both lack explicit psychometric grounding. In contrast, our dataset and framework are the first to study the causal link between latent Big Five personality traits and downstream preferences, shifting the evaluation focus from simple rule adherence or fact recall to psychologically plausible, trait-driven preference prediction.

4.2.7 Conclusion

In conclusion, we introduce PACIFIC, a psychometrics-based framework demonstrating that leveraging latent personality traits enhances LLM preference alignment accuracy, improving label-free retrieval accuracy by 43% relative (30→43%) over standard semantic retrieval, while clean trait-aligned context approaches ceiling.

4.3 Discussion

For personalization, memory should be a compact semantic abstraction of the person, not a lossless replay of their history. Given clean, trait-aligned context, models reason about who a person is almost perfectly. Assembling that context is hard, and more raw history

actively hurts: mismatched preferences add noise and also drag the answer away from the person. The bottleneck of personalization therefore lies in the representation of the person, not in the model's reasoning. A compact, stable trait profile that compresses history is the representation this chapter uses. Enlarging the context window does not solve the problem.

This chapter also surfaces a caution that connects personalization back to fairness. When personality is inferred from behavior, the inference is not evenhanded across the trait spectrum: "high" expressions of a trait are recovered far more reliably than "low" ones, because alignment training has learned to treat the socially desirable pole (outgoing, open, reassuring) as the default and to read the opposite pole as something to be corrected. For a large share of the trait space (introverted, routine-preferring, risk-tolerant users), personalization can silently substitute the normatively approved answer for the person's actual one, without signaling that it has. A system built to adapt to individuals can therefore reintroduce, at the level of persona, the unfairness that Chapters 2 and 3 worked to remove at the level of the group. Faithful personalization is a fairness requirement as well.

This chapter has personalized to one person through a stable model of who they are. An individual is not the only unit a generative model must serve: it is deployed for many communities at once, whose values genuinely differ and sometimes conflict. Learning those group preferences also raises the privacy and equity concerns of the earlier chapters, since preference data is sensitive and no group should be aligned away in favor of a majority. Chapter 5 takes up alignment to many groups at once, privately and fairly.

Chapter 5

Pluralistic Alignment of Large Language Models

5.1 Overview

Chapter 4 personalized a generative model to a single individual. This chapter considers the collective case. A deployed language model serves many communities at once, whose values differ across region, age, and culture and sometimes directly conflict. Aligning a model to this plurality cannot be done by aligning to an "average" person: the prevailing alignment method, reinforcement learning from human feedback (RLHF), optimizes toward a single aggregated preference and, as a result, reinforces majority viewpoints while marginalizing minority ones. *Pluralistic alignment* is the goal of building models that represent diverse human values faithfully, without collapsing them.

Pursuing that goal re-engages the concerns of the earlier chapters. Learning group preferences from human feedback means collecting and centralizing sensitive preference data, the kind of exposure Chapter 3 studied, and it raises again the question of whose preferences the

aligned model prioritizes and whose it leaves behind. This chapter therefore treats pluralistic alignment as a federated, privacy-preserving, and fairness-aware problem, in two papers that build on one another.

The first, **PluralLLM** [Srewa et al., 2025a], asks how to *learn* group preferences without centralizing them. It trains a lightweight transformer-based preference predictor across user groups using federated learning, so that each group contributes to a shared model without exposing its raw preference data. The predictor captures group-specific preference patterns, generalizes to unseen groups, and can serve as a lightweight reward model in place of the expensive reward-modeling stage of conventional RLHF. The second, a **systematic evaluation of preference aggregation in federated RLHF** [Srewa et al., 2025b], asks how those learned group preferences should be *combined*. In a federated RLHF pipeline, each group evaluates the policy's rollouts locally and returns a group-specific reward, and the server must aggregate these heterogeneous, sometimes conflicting reward signals into the single reward that drives policy optimization. That aggregation choice determines whose preferences are honored. [Srewa et al., 2025b] compares standard rules (min, max, average) against a proposed *adaptive* scheme that dynamically upweights groups with weaker historical alignment, across a range of reward functions and both probability-prediction and ranking formulations of the task.

Section 5.2 presents PluralLLM and Section 5.3 presents the aggregation study. Section 5.4 discusses both.

5.2 PluralLLM: Pluralistic Alignment in LLMs via Federated Learning

5.2.1 Introduction

Large Language Models (LLMs) have rapidly emerged as a cornerstone of modern artificial intelligence, powering applications ranging from conversational agents to content generation and decision support systems Bubeck et al. [2023]. Their ability to generate human-like text has revolutionized various industries, but their effectiveness depends on their ability to align with human values and societal expectations Zhang et al. [2023]. However, achieving robust human alignment remains a significant challenge, particularly in ensuring that these models fairly represent diverse perspectives, a concept known as Pluralistic Alignment Sorensen et al. [2024].

Existing LLM alignment methods fall into two categories: prompt engineering and gradient-based alignment Sahoo et al. [2024], Zhang et al. [2023]. While prompt engineering guides model behavior through carefully crafted prompts and in-context examples without modifying model parameters, it often struggles with complex behaviors Zhou et al. [2022]. Gradient-based alignment fine-tunes models using reward mechanisms, such as Reinforcement Learning from Human Feedback (RLHF), improving traits like honesty and helpfulness but requiring extensive supervision and high computational costs Ouyang et al. [2022]. However, existing approaches do not scale efficiently for aligning LLMs across multiple user groups with limited supervision, making pluralistic alignment challenging Chakraborty et al. [2024], Casper et al. [2023].

Despite recent advancements in alignment techniques, preference learning for Large Language Model (LLM) faces three significant challenges: (1) privacy risks, (2) collection of preference data, and (3) computational overhead Kuang et al. [2024], Casper et al. [2023]. Centralized approaches, such as RLHF and gradient-based alignment, require collecting and

processing user interactions, raising concerns about data security and user confidentiality. On the other hand, FL enables privacy-preserving model training by keeping data decentralized. However, it comes with high communication and processing costs due to frequent updates between clients and global server Kuang et al. [2024] . These challenges are even greater for pre-trained LLM, which require significant computational resources to incorporate preference data. Aligning LLM with FL adds further complexity, demanding a balance between alignment quality, efficiency, and privacy. In preference learning, this burden is particularly heavy, making efficient aggregation strategies or alternative models essential Wu et al. [2024]. In PluralLLM, we address the following research question:

How can we align LLMs to capture the preference of various perspectives of different communities while preserving privacy, maintaining fairness, and ensuring computational efficiency?

We introduce **PluralLLM**, a framework for pluralistic alignment in LLMs via FL. Our approach leverages FL to train a transformer-based preference predictor Zhao et al. [2023a] to capture group-specific preferences in a distributed privacy-preserving manner. This preference predictor serves as a lightweight alternative to conventional alignment methods that require training a reward model RLHF, significantly reducing computational overhead and can adapt to a new unseen group. Unlike the centralized group preference optimization training approach proposed by Zhao et al. [2023a], our FL-based preference predictor ensures privacy-preserving preference learning by allowing different groups to collaboratively train the model without exposing their sensitive preference data. In addition, it enables diverse groups to participate in training while maintaining the fairness properties of the centralized approach. Our results demonstrate that our proposed **PluralLLM** approach applied to preference learning achieves higher alignment scores and faster convergence compared to centralized methods, making it a scalable and efficient solution for capturing diverse group preferences.

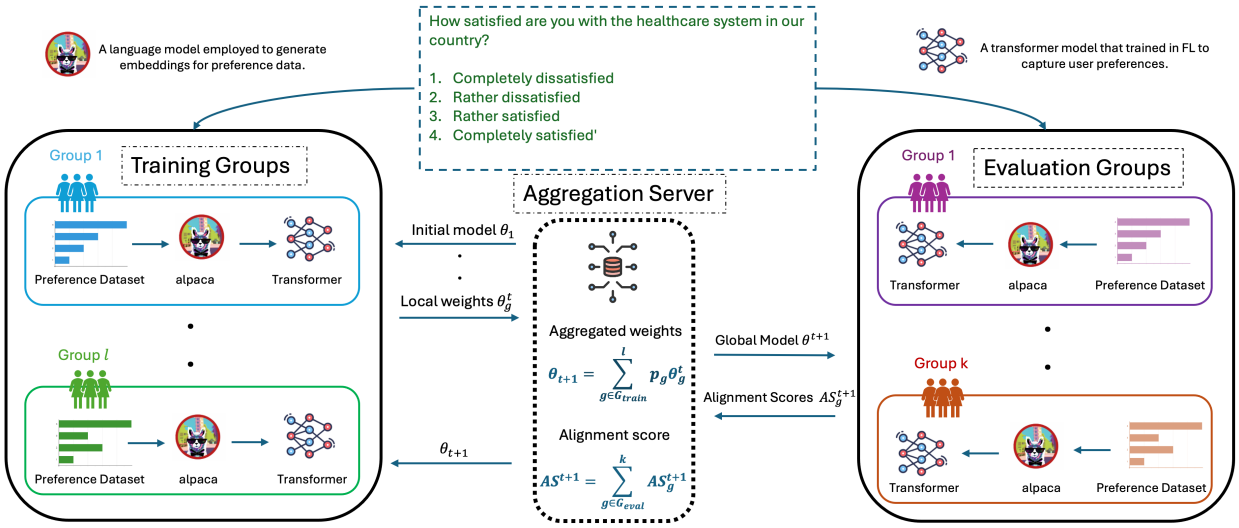


Figure 5.1: **PluralLLM**: Pluralistic alignment in LLMs via Federated Learning.

5.2.2 Related Work

Prompt Engineering: Prompt engineering provides a mechanism for fine-tuning model outputs through the modification of the input to the LLM, thereby aligning with user preferences without altering the parameters of the core LLM model. Prompt engineering approaches are characterized by their computational efficiency, a property that stems from the absence of any training requirements Sahoo et al. [2024]. However, the design of the prompt itself can be a laborious task that relies on heuristics. The efficacy of these heuristics is not guaranteed to transfer well across different LLMs Zhao et al. [2023a]. Recent work in the literature has shown that prompt engineering has limited success in aligning LLMs to complex groups on challenging survey datasets such as *GlobalOpinionQA* Durmus et al. [2023].

Pluralistic Alignment in LLM: A growing number of pluralistic alignment studies show that it is important to design LLM systems that can accommodate and represent diverse human values, perspectives, and preferences. Unlike traditional alignment approaches that aim to align models to a single, averaged set of human preferences, pluralistic alignment seeks to reflect the complexity and plurality of human societies. For example, Cao et. al. introduced an age fairness reward in LLM to reduce response quality disparities across

distinct age groups during training Cao et al. [2024]. Traditional Reinforcement Learning alignment approaches, such as RLHF often reinforce majority viewpoints while marginalizing minority perspectives. The question of balancing openness to diverse values with ethical constraints, such as the avoidance of harmful ideologies, remains largely unaddressed.

Group Preference Alignment: Group preference alignment refers to techniques designed to adapt LLM outputs to reflect the distinct preferences, values, or judgments of different groups or demographics. Group Preference Optimization (GPO) Zhao et al. [2023a] was introduced as a few-shot alignment framework that steers LLMs toward group-specific preferences. GPO augments the base LLM with an independent transformer module, trained via in-context supervised learning with only a handful of samples to predict group preferences and refine model outputs. This module acts as a preference model for different groups, learning distinct alignment patterns across diverse communities. By leveraging an in-context autoregressive transformer, GPO enables flexible and efficient alignment, allowing LLMs to adapt dynamically to varying user preferences.

5.2.3 Pluralistic Alignment in LLMs via Federated Learning

We chose the Q/A preference alignment task, which involves aligning LLM responses based on group-specific preferences. This task is particularly well-suited for evaluating pluralistic alignment, as it requires the model to adapt to diverse user opinions while maintaining coherence and fairness. Unlike standard classification tasks, preference-based Q/A alignment provides a richer evaluation metric, allowing us to measure not only the correctness of responses but also how well the model captures nuanced group preferences. This setup also reflects real-world applications, where LLM must personalize responses based on collective user preferences. We adopt a FL setup for pluralistic alignment in LLM. The framework consists of three main actors, as described in Figure 5.1:

- **Training Clients (Groups):** The training set, G_{train}, comprises l distinct groups, each

representing a client in a FL setup. Each client performs local training to develop a transformer-based preference model Zhao et al. [2023a]. This model aims to learn group-specific preference patterns and generalize to unseen data. Each client trains its transformer independently using its respective group's preference dataset.

- **Aggregation Server:** A central server coordinates FL by collecting, aggregating, and redistributing model updates from training clients. This process enables learning from diverse data while preserving privacy and preventing direct data sharing.

- **Evaluation Clients (Groups):** A separate set G_{eval} consisting of K groups is introduced, where each group acts as a client in the FL setup to assess the alignment performance of the trained model. Unlike training clients, these groups do not participate in model updates. Instead, they represent new unseen groups that serve as an independent benchmark to evaluate the generalizability of the trained model. Their feedback helps determine how well the aggregated model aligns with unseen groups.

5.2.3.1 Group Local Training

Each training group $g \in G_{\text{train}}$ has a preference dataset: $D_g = \{(x_1^g, y_1^g), \ldots, (x_n^g, y_n^g)\}$ where x_i^g represents the embedding of LLM concatenated prompt-response pair: $x_i^g = \omega_{\text{emb}}(q_i^g, r_i^g)$ and ω_{emb} is a language model embedding function. The preference y_i^g represents the group preference probability for the generated response r_i^g to the query q_i^g. For instance, in Figure 5.1, the query represents the question, while the response corresponds to the aggregated probability distribution of answers within a group. This preference data is then processed by the Alpaca model (LLM) to generate embeddings, following the approach in Zhao et al. [2023a].

The dataset for each group is randomly divided into m context samples $(x_1^g, y_1^g), ..., (x_m^g, y_m^g)$ and $n-m$ target samples $(x_{m+1}^g, y_{m+1}^g), .., (x_n^g, y_n^g)$. The model is trained to predict the target preferences given the context examples, optimizing the following loss function as introduced

in Zhao et al. [2023a]:

$$\mathcal{L}(\theta) = \mathbb{E}_{g,m}\left[\sum_{i=m+1}^{n} \log p_\theta(y_i^g \mid x_{1:m}^g, y_{1:m}^g, x_i^g)\right], \tag{5.1}$$

where p_θ denotes the target points predicted preference distribution conditioned on the context points. At the end of local training, each client g transmits its updated model parameters θ_g to the central aggregation server, which combines the received updates to train the global model.

5.2.3.2 Model Aggregation

We employ the FedAvg technique to aggregate updates from multiple groups McMahan et al. [2017]. The aggregation process is designed to minimize the global optimization function of FL.

$$\min_\theta \; F(\theta) = \sum_{g \in G_{\text{train}}} p_g F_g(\theta), \tag{5.2}$$

where p_g is the weight assigned to each training group g, defined as: $p^g = \frac{|D_g|}{\sum_{g'} |D_{g'}|}$. Here, D_g represents the size of the preference dataset for group g. The term p_g ensures that each group's contribution to the global objective is weighted by the proportion of its dataset size relative to the total dataset across all training groups. The local objective function for group g is defined as: $F_g(\theta) = \mathbb{E}_{(x,y)\sim D_g}\left[\mathcal{L}(f_\theta(x), y)\right]$, where $\mathcal{L}$ represents the loss function defined at Equation 5.1 and $f_\theta(x)$ is the model output for input x. To approximate the global objective, model updates from training groups are aggregated as follows:

$$\theta^{t+1} = \sum_{g \in G_{\text{train}}} p_g \theta_g^t, \tag{5.3}$$

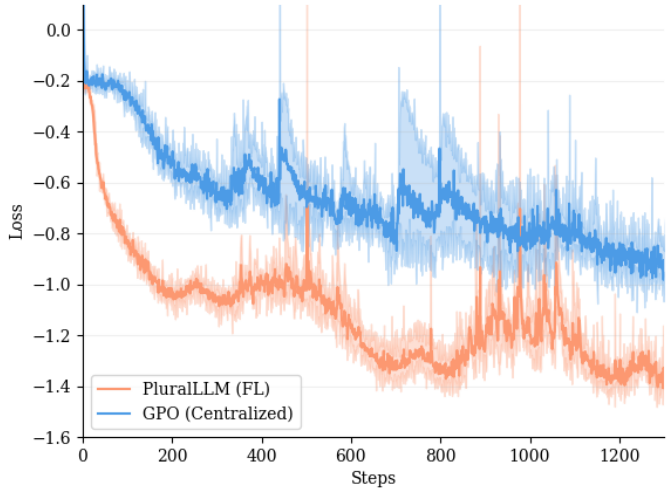


Figure 5.2: Comparison of training loss curves for centralized learning GPO and **PluralLLM**. **PluralLLM** achieves a lower loss compared to Centralized Training GPO.

where θ_g^t represents the locally updated model parameters at the training group g in round t. The aggregated model is then redistributed to training and evaluation clients.

5.2.4 Experiments

5.2.4.1 Experimental Setup

Our experiments utilize a transformer-based preference predictor model (GPO Zhao et al. [2023a]), originally designed to train groups sequentially in an ordered manner. However, in **PluralLLM**, we adapt it to a FL setting, employing FedAvg as a completely different learning paradigm. The primary goal of our evaluation is to determine whether our approach in **PluralLLM** impacts the alignment score and group fairness across different groups compared to the original centralized learning.

Experiments are conducted on machines equipped with one NVIDIA A30 GPU, an Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz, and 256GB RAM. All results reported in this study are averaged over four different runs with varying random seeds to ensure stability.

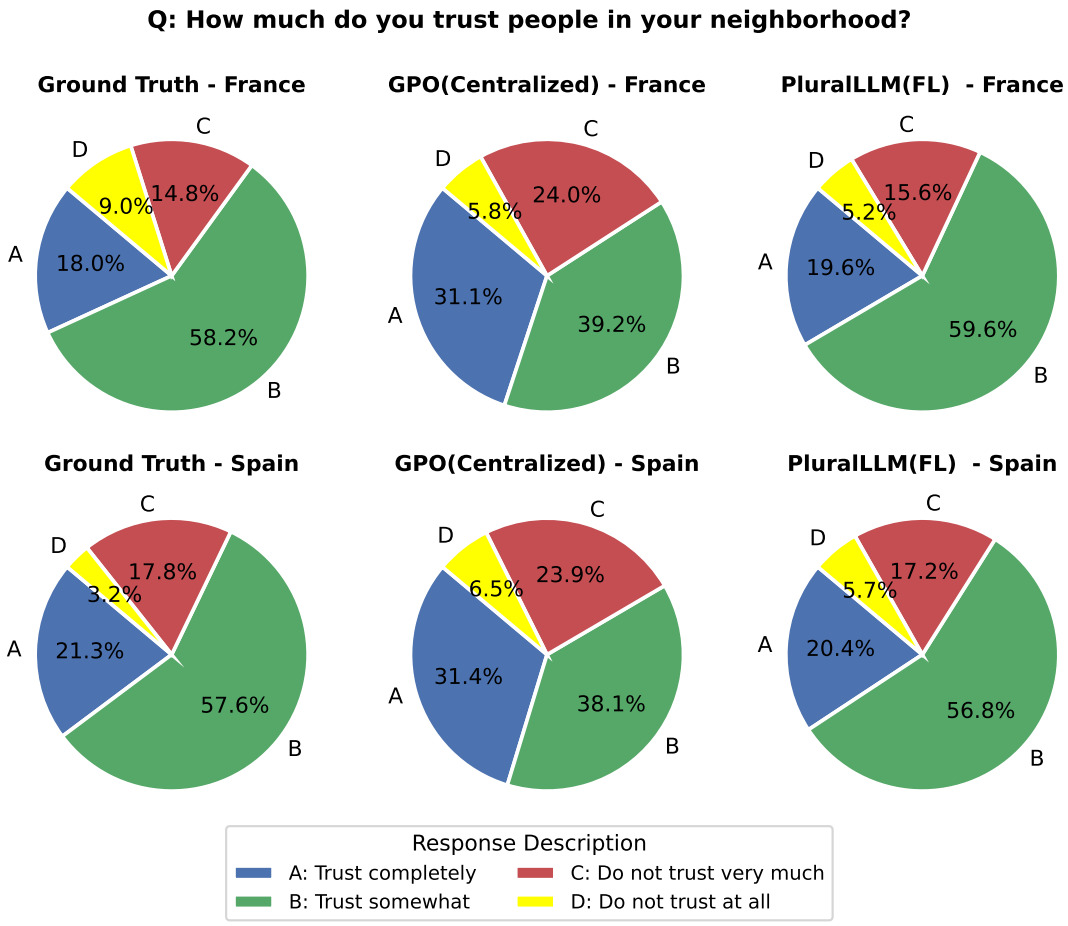


Figure 5.3: Comparison of preference distributions across Ground Truth, Centralized Learning GPO, and **PluralLLM** for a given question.

5.2.4.2 Dataset

We used the dataset from the Pew Research Center's Global Attitudes Surveys (PewResearch), which collects public opinions on a wide range of social, political, and economic issues Durmus et al. [2023]. These surveys capture diverse perspectives from various demographic and geographic groups, providing a rich foundation for analyzing preference alignment. To ensure a fair comparison with GPO Zhao et al. [2023a], we use the same subset of groups (13 in total) as in the original GPO training setup. In both FL and centralized learning, groups are partitioned into 60% and 40% groups for training and evaluation, respectively.

5.2.4.3 Implementation

- **Federated Learning:** We trained the transformer-based preference predictor model (GPO) for $1,300$ communication rounds in the FL setup, assuming that all clients participate in each communication round. Each local training step consists of 6 local epochs, where, in each epoch, we randomly sample context questions and corresponding preferences, then the target questions that we wish to predict its preferences to train the model. Adam optimizer used with a learning rate of 3×10^{-4}.

- **Centralized Learning:** We trained the transformer-based preference predictor model (GPO) for $1,300$ epochs, iterating over all training groups in each epoch. During the training epoch, each group samples a random selection of context questions and their corresponding preferences and target questions. Unlike FL, where model updates are aggregated after each communication round, the centralized approach updates the model sequentially for each group within a single epoch.

- **Preferences Embedding**: We use Alpaca-7B, a fine-tuned version of LLaMA-7B, as the embedding model to represent preference data, which is then passed to transformer input Zhao et al. [2023a]. The embedding step is done once over all the preference data for each group at the beginning of training.

To assess alignment performance, evaluation is conducted every 10 communication rounds (in the FL setting) or every 10 epochs (in the centralized setting). The alignment score is computed over randomly sampled data from all the evaluation groups to measure how well the trained model adapts to new group preferences over time.

5.2.4.4 Evaluation Metrics

To quantify the impact of **PluralLLM** on alignment performance, we evaluate:

- **Alignment Score(`AS`):** We assess the degree of alignment between 2 opinion distributions

P_1 and P_2 by calculating the Alignment Score $AS(P_1, P_2; Q)$ over a set of questions Q as used in Zhao et al. [2023a]. Jensen-Shannon Distance [1], denoted as *JSD*, is used to assess preference distribution similarity shifts.

$$AS(P_1, P_2; Q) = \frac{1}{|Q|} \sum_{q \in Q} JSD(P_1(q), P_2(q); Q) \tag{5.4}$$

- **Convergence Speed of Loss Function:** We measure how quickly the model optimizes preference alignment by tracking the average loss across all clients at each communication round. This is compared to the centralized training loss per epoch. Convergence speed is defined as the point where the model reaches 95% of its final loss value.

- **Fairness Metrics:** We analyze the effect of **PluralLLM** on group fairness in preference alignment across different groups compared to centralized learning. Numerous studies have demonstrated that FL can inadvertently introduce unfairness into the trained models Shi et al. [2023]. This unfairness arises primarily from data heterogeneity across clients, which leads to disparate performance results and challenges in achieving equitable model accuracy across all participants. Furthermore, these fairness issues have the potential to show disparity in privacy leakage risks, as adversaries can potentially exploit shared model parameters to infer sensitive information Zhao et al. [2026b]. We assess the fairness by adapting Coefficient of Variation (CoV) and Fairness Index(FI) to measure the disparity of alignment scores across distinct groups. These metrics are used to measure the relative perception of fairness in human-centered systems Zhao and Elmalaki [2024], Zhao et al. [2024]. For K groups, we define the alignment score of group i as AS_i. The average alignment score across groups is calculated as $\mu = \frac{1}{K} \sum_{i=1}^{K} \mathrm{AS}_i$. The CoV of alignment scores $CoV(AS)$ is calculated as:

$$CoV(AS) = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{K} \sum_{i=1}^{K} (\mathrm{AS}_i - \mu)^2}}{\mu}, \tag{5.5}$$

[1]The Jensen–Shannon divergence (JSD) is a symmetric measure of similarity between two probability distributions, always non-negative, with 0 denoting identical distributions and any value above 0 indicating differences.

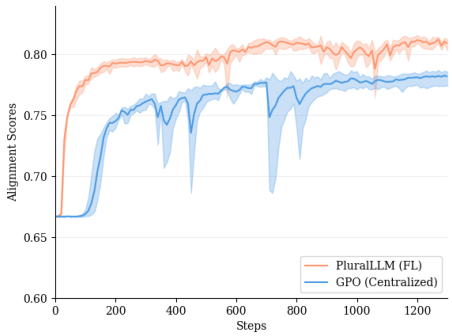


Figure 5.4: Comparison of mean evaluation group alignment scores for centralized learning GPO and **PluralLLM**.

where σ is the standard derivation of AS. A lower CoV indicates a more equitable distribution of alignment scores among groups, suggesting better fairness in aligning distinct groups. We apply the Fairness Index (FI) transformation to interpret the fairness in percentage between 0 and 1. A higher FI indicates greater fairness, where 1 represents perfect fairness.

$$FI(AS) = \frac{1}{1 + CoV^2(AS)} \tag{5.6}$$

5.2.4.5 Analysis of Convergence Speed

PluralLLM converges at communication round 634, whereas the centralized approach requires significantly more steps, converging at iteration 1180 epoch as seen in Figure 5.2. Hence, **PluralLLM** achieves convergence **46%** faster than the centralized learning approach, highlighting its efficiency in accelerating model training. Additionally, **PluralLLM** maintains a lower loss throughout training as observed in Figure 5.2, demonstrating improved stability compared to the centralized approach. The faster convergence of **PluralLLM** suggests that it is well-suited for distributed learning scenarios where reducing communication rounds is critical for efficiency.

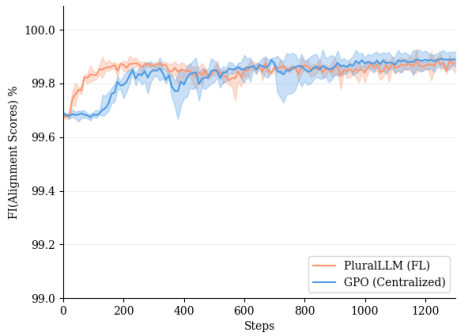


Figure 5.5: Comparison of Fairness Index between centralized learning and **PluralLLM**. The utilization of **PluralLLM** in the training of a preference transformer *Does Not* result in significant disparities among groups.

5.2.4.6 Analysis of Alignment Performance

Figure 5.4 demonstrates that **PluralLLM** achieves a $\approx 4\%$ improvement in the mean evaluation alignment score compared to the centralized approach. While the centralized method shows slower improvements and remains at a lower score, **PluralLLM** maintains a more stable progression with lower fluctuations, indicating that the distributed training approach enhances robustness in achieving better alignment and improves generalization across diverse data distributions. Additionally, Figure 5.3 highlights that for two evaluation groups, **PluralLLM** more accurately represents the baseline distribution for a given Q/A task compared to the centralized approach.

5.2.4.7 Analysis of Fairness in Alignment

As observed in Figure 5.5, **PluralLLM** improves the FI by 0.04% on average before converging at a round 634 compared to the centralized GPO. **PluralLLM** maintains a comparable FI across training steps till round 1300 achieving an equal opportunity with $FI \approx 1$. The higher FI of **PluralLLM** in the early rounds is consistent with its faster convergence under federated training.

5.2.5 Discussion & Conclusion

PluralLLM introduces a federated learning-based approach for pluralistic alignment in LLMs, addressing privacy, efficiency, and scalability challenges. Decentralizing the training of a transformer-based preference predictor preserves user privacy while capturing diverse group preferences more effectively than centralized methods. Our evaluation employs FedAvg for efficient preference update aggregation, resulting in 46% faster convergence, a 4% improvement in alignment scores, and maintaining group fairness comparable to centralized training.

In addition, this predictor can serve as a lightweight reward function for RLHF, reducing computational costs or generating high-quality preference datasets for DPO, improving efficiency. While effective in Q/A tasks, its applicability to other domains like summarization and translation remains unexplored. Future work will focus on integrating learned preferences into LLM fine-tuning methods and exploring alternative aggregation strategies to enhance fairness across tasks. Furthermore, extending **PluralLLM** beyond Q/A to diverse learning tasks.

5.3 A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs

5.3.1 Introduction

The remarkable capabilities of LLMs have positioned them as a central technology across various domains. However, their real-world utility and safety hinge on their ability to align with complex and diverse human values and social normsYang et al. [2024], Sorensen et al. [2024]. The prevailing methodology for this alignment is RLHF, which fine-tunes models based on collected human preference data Ouyang et al. [2022]. While effective, the standard

RLHF paradigm often operates on a centralized dataset, which is not only a privacy concern but also risks embedding biases of a narrow demographic Casper et al. [2023].

To address this, the integration of RLHF with FL has emerged as a promising avenue. FL allows for model training on decentralized data from numerous clients, thus preserving data privacy and capturing a wider range of human preferences Wu et al. [2024], Srewa et al. [2025a]. However, this fusion presents a critical and underexplored challenge: **How to aggregate the diverse and potentially conflicting preference signals from different user groups?** In our setting, these preference signals appear as per-group reward scores generated locally by each client; the server aggregates these decentralized reward vectors without accessing any raw data to compute a global reward used for PPO updates. The choice of aggregation strategy is not merely a technical detail; it is an evaluation protocol that directly shapes the model's final behavior, determining whose preferences are prioritized and whose are marginalized.

This evaluation proposes a systematic evaluation framework to analyze the impact of different aggregation techniques on both alignment performance and fairness. By comparing standard reward aggregation methods with our proposed adaptive aggregation scheme, our goal is to define a more robust protocol to assess LLMs in decentralized, pluralistic environments. We show that while simple aggregation methods can lead to unintended biases, our adaptive approach strikes a superior balance between achieving strong overall alignment and ensuring equitable representation across diverse groups, thus contributing to the development of more reliable and justly aligned LLMs. Our approach follows a zero-shot alignment paradigm, using only aggregated group reward signals without task demonstrations, ensuring generalizable alignment.

5.3.2 Background and Related Work

The alignment of LLMs with complex human preferences is a central goal in their development. Because explicitly encoding human values into a fixed loss function is challenging, *Reinforcement Learning from Human Feedback* (RLHF) has emerged as the dominant paradigm for steering models toward desired behavior. In RLHF, models learn from human preference data via a reward model trained on pairwise comparisons, and the policy is then optimized to maximize this learned reward Ouyang et al. [2022]. The most commonly used RL algorithm in RLHF is Proximal Policy Optimization (PPO), which fine-tunes the LLM using feedback from the reward model trained on human preference pairs Schulman et al. [2017]. An alternative, Direct Preference Optimization (DPO), simplifies this pipeline by bypassing an explicit reward model and directly optimizing the policy to assign higher probability to preferred responses Rafailov et al. [2023].

While traditional alignment methods typically aim for a single, global preference objective, real-world users exhibit diverse and sometimes conflicting preferences. Group Preference Optimization (GPO) Zhao et al. [2023a] was introduced to address this challenge by enabling group-specific alignment of LLMs. GPO augments the base LLM with a lightweight transformer module trained via in-context supervised learning to predict and incorporate the distinct preferences of user groups using only a few examples. This auxiliary module serves as a preference predictor that captures alignment patterns across heterogeneous communities, allowing the model to dynamically adapt its responses according to different social or demographic contexts. In parallel, methods such as Group Robust Policy Optimization (GRPO) Ramesh et al. [2024] and MaxMin-RLHF Chakraborty et al. [2024] have focused on ensuring robustness across diverse user populations within centralized RLHF settings. However, these methods still rely on collecting and processing user data on a central server, which raises privacy and data ownership concerns. To overcome these limitations,

To overcome these limitations, PluralLLM Srewa et al. [2025a] extends GPO Zhao et al.

[2023a] into a federated learning architecture that enables groups to collaboratively learn lightweight transformer-based preference predictors without sharing raw data. Concretely, each group trains a lightweight transformer using few-shot in-context examples in a FL manner, producing a local preference module that can predict, for any Q/A question, a probability distribution over all answer options reflecting that group's latent preferences. This module serves as a fully local reward model; given RLHF rollouts from the server, the PluralLLM outputs group-specific preference probabilities that can be transformed into scalar rewards.

Our work builds directly on this foundation by using these PluralLLM predictors as decentralized reward generators for each group. At each PPO iteration, the server broadcasts rollouts to all groups, each group evaluates the responses using its PluralLLM module, and returns group-specific reward vectors. The central question we address is how to aggregate these heterogeneous and sometimes conflicting reward signals across groups. By systematically comparing standard aggregation schemes and introducing a dynamic alpha aggregation strategy, we analyze how different reward-aggregation protocols affect both fairness and alignment quality in multi-group federated RLHF.

5.3.3 Methodology

System Setup and Training Groups: In our setting, each group g_i corresponds to a distinct demographic or preference cluster (e.g., age, region, political leaning), and each group acts as a single federated client that locally represents its users' aggregated preferences. Our framework focuses on Q/A tasks with l training groups $G_{\text{train}} = \{g_1, g_2, \ldots, g_l\}$, where each group g_i maintains its private preference dataset $D_{g_i} = \{(x_{i,j}, y_{i,j})\}$ locally. Each preference sample consists of a query-response pair embedding $x_{i,j}$ and the corresponding group preference probability $y_{i,j}$. These datasets are distributed across groups and never shared with the central server, ensuring privacy preservation.

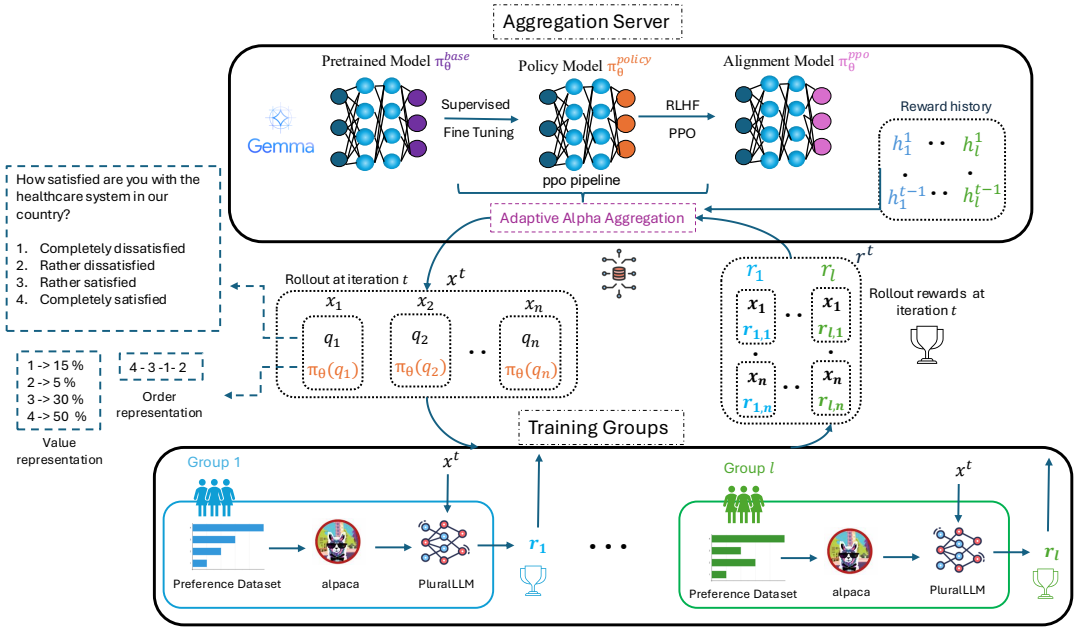


Figure 5.6: Federated RLHF for pluralistic alignment of group preferences in LLM.

As illustrated in Figure 5.6, the aggregation server is initialized with a base LLM model π_θ^{base} and performs supervised fine-tuning (SFT) to adapt it for Q/A tasks, resulting in a policy model π_θ^{policy} suitable for PPO training. The server coordinates between policy optimization and distributed preference learning.

At iteration t, the server generates rollouts X^t consisting of queries (questions with multiple choice's options) and LLM responses using the current policy π_θ^{policy}. These rollouts are distributed to all training groups for preference evaluation.

Distributed Reward Generation: Each group g_i first uses the PluraLLM Srewa et al. [2025a] as a local lightweight reward model to generate preference probabilities for the received rollouts at iteration t. These probabilities are then converted to rewards $r_{g_i}^t$ according to the reward metric used. In our evaluation, we focus on two approaches: (1) preference probability prediction, where rewards are calculated directly from the predicted probabilities, and (2) preference ranking, where the probabilities are first converted to rankings before reward calculation. These task-specific rewards $r^t = \{r_{g_1}^t, r_{g_2}^t, \ldots, r_{g_l}^t\}$ reflecting how

well the generated responses align with each group's preferences, are transmitted back to the aggregation server.

Adaptive Alpha Aggregation: The core of the federated RLHF framework is the aggregation of local rewards. At each training round, the central server receives per-group rewards r^t from different groups and aggregates them into a final global reward $Agg_\alpha(\mathbf{r}^t)$, ready to be used for updating the policy. We have $r^t = \{r^t_{g_1}, r^t_{g_2}, r^t_{g_3}, \ldots, r^t_{g_l}\}$ from l clients, where r^t_i represents how well the LLM outputs align with group g_i preference at FL iteration t. Recent work in the literature introduced an aggregation method, namely alpha aggregationPark et al. [2024] which achieves the consensus among heterogeneous feedback in RLHF. This is highlighted in Equation 5.7. This consensus reward aggregation is controlled by α, $Agg_\alpha(\mathbf{r}) = max(\mathbf{r})$, when $\alpha = \infty$ and $Agg_\alpha(\mathbf{r}) = min(\mathbf{r})$, when $\alpha = -\infty$.

$$Agg_\alpha(\mathbf{r}) = \begin{cases} \frac{1}{\alpha}\log(\frac{1}{N}\sum_{i\in N}\exp(\alpha r_i)) & \alpha \neq 0 \\ \frac{1}{N}\sum_{i\in N} r_i & \alpha = 0 \end{cases} \tag{5.7}$$

We propose an *adaptive* extension of this scheme that replaces the single global α with *group-specific*, dynamically updated weights α^t_g. Instead of using one fixed α for all clients, our method learns α^t_g per group based on its historical alignment performance, and plugs these into the alpha aggregation:

$$Agg_\alpha(\mathbf{r}^t) = \begin{cases} \frac{1}{|G_{\text{train}}|}\sum_{g\in G_{\text{train}}} r^t_g & \text{if } FI \geq 0.9 \\ \log\left(\frac{1}{|G_{\text{train}}|}\sum_{g\in G_{\text{train}}}\exp(\alpha^t_g \cdot r^t_g)\right) & \text{otherwise} \end{cases} \tag{5.8}$$

To achieve a balanced accumulated alignment reward history $\mathbf{h} = \{h_{g_1}, h_{g_2}, h_{g_3}, \ldots, h_{g_l}\}$, across clients, the aggregation weights α_i are dynamically adjusted in inverse proportion to each client's historical alignment performance($\alpha_i = \texttt{softmax}(1 - h_i)$). Specifically, a client i

with a lower accumulated alignment reward, h_i, is assigned a higher weight, α_i. As shown in Equation 5.8, a higher α_i value increases the dominance of the corresponding reward r_i. Hence, the α_i value for client i changes adaptively based on the alignment history for this client h_i.

Fairness Index (FI): To determine when uniform averaging is sufficient versus when adaptive weighting is needed, we use a FI that measures the similarity of per-group rewards $r^t_{g_i}$ across all groups. FI ranges from 0 to 1, where values near 1 indicate highly consistent (fair) rewards and lower values indicate increasing disparity. When $FI \approx 1$ (we use a practical threshold $FI \geq 0.9$), the rewards across groups are nearly uniform, and thus a simple average aggregation is appropriate. Otherwise, we employ the α-weighted log-sum-exp aggregation to amplify contributions from groups with historically lower alignment performance.

PPO Training and Iteration: With the aggregated rewards $Agg_\alpha(r^t)$, the server performs PPO optimization to update the policy model $\pi^{policy}_{\theta^t} \rightarrow \pi^{policy}_{\theta^{t+1}}$. The updated policy generates new rollouts for the next iteration, and the process continues until a predefined number of iterations or specific alignment score is reached. This iterative approach ensures continuous adaptation to diverse group preferences while maintaining fairness through our adaptive aggregation scheme.

5.3.4 Evaluation

We evaluate our approach using Gemma-2B-it, a fine-tuned version of the Gemma model, as our base LLM Team et al. [2024], More details on the experiment setup and configuration are summarized in Appendix D.1.

Our experiments utilize the Pew Research Center's Global Attitudes Surveys dataset Durmus et al. [2023], which captures diverse public opinions across social, political, and economic

issues from various demographic groups. The dataset consists of multiple-choice survey questions answered by participants from a wide range of countries, with 2,554 questions spanning topics such as politics, media, technology, religion, race, and ethnicity. For each question and country, the dataset provides a probability vector over the answer choices, indicating the fraction of respondents in that country selecting each option. These probability vectors differ across countries, reflecting distinct group-level preferences. In our experiments, we treat each country as a separate user group (i.e., a federated client) and use all available groups in the survey. Our goal is to align the LLM with these diverse group preferences in a fair and robust manner, without overfitting to any single group or majority and without leaving any group behind.

We assess performance using fairness index FI and alignment scores across two primary tasks: the *preference probability prediction task* (see Figure 5.7) and the *preference ranking task* (see Figure 5.8). Our evaluation framework encompasses various reward functions and aggregation strategies. We compare our adaptive alpha aggregation against standard federated approaches (Min, Max, Average) and a supervised fine-tuning (SFT) baseline.

5.3.4.1 Evaluation Framework

The LLM rollout at FL iteration t, denoted X^t, consists of a set of questions $\{q_j\}$ together with corresponding responses generated by the policy model π_θ for each question q_j. We parse each response to extract the relevant output and denote the resulting LLM prediction as $y_j^{t,\text{llm}}$ for question j at iteration t. Let $o_{j,k}$ be the k-th answer option for question j, and let K denote the total number of answer options for that question. For each group $g \in G_{\text{train}}$, we compute a reward $r_{g,j}$ by comparing the LLM prediction $y_j^{t,\text{llm}}$ against the PluralLLM-derived target distribution for that group, $p_{g,j}^{\text{PluralLLM}}$ Srewa et al. [2025a], i.e., the group-specific probability vector over options for question j predicted by the local PluralLLM reward model.

Preference Probability Prediction Prompt

<bos><start_of_turn>user

You are an expert in modelling group preferences. You will receive **a question and exactly 4 options**.

Your task

- Assign a preference score to **each and every option**
- Produce **4** scores—no option may be skipped or combined
- Each score must be a decimal between 0 and 1, and **the rounded scores must sum to 1.00**
- Higher scores represent options a typical group is more likely to choose

Output format

- One line, comma-separated decimal numbers, **no spaces**
- Round each to **2** decimal places
- **No extra text, labels, or symbols**
- Example: 0.65,0.20,0.10,0.05

Return ONLY the 4 scores in the same order as options.

Question: Germany's influence in the EU Options: A: Has too much influence B: Has too little influence C: Has about the right amount of influence D: DK/Refused

<end_of_turn> <start_of_turn>model

Figure 5.7: Preference Probability Prediction Prompt

Preference Ranking Prompt

<bos><start_of_turn>user

You are an expert in ranking group preferences. You will receive **a question and exactly 4 options**.

Your task

- Rank all **4 provided options** from most to least preferred
- Process every option—no skipping or combining
- Order options based on what a typical group would most likely choose
- Higher preference options appear first

Output format

- One line, comma-separated option letters, **no spaces**
- Use the exact provided letters
- **No extra text, labels, or symbols**
- Example: B,C,A,D

Return ONLY the 4-letter ranking.

Question: Germany's influence in the EU Options: A: Has too much influence B: Has too little influence C: Has about the right amount of influence D: DK/Refused

<end_of_turn> <start_of_turn>model

Figure 5.8: Preference Ranking Prompt

5.3.4.2 Reward Metrics

Our evaluation employs two categories of reward metrics to assess alignment quality across different aspects of preference modeling. These rewards are also used as alignment scores in our evaluation, i.e., they directly define the Avg AS and Min AS reported in Table 5.1.

Distance-Based Reward Metrics (Preference Prediction Task) These rewards quantify the alignment between the LLM-predicted and PluralLLM-target probability distributions. They are computed locally on each client in G_{train}, where each group $g \in G_{\text{train}}$ independently evaluates the following reward functions:

Wasserstein Reward: Measures optimal transport cost between distributions.

$$r_{g,j}^{t,\text{Was}} = \frac{W_1(y_j^{t,\text{llm}}, p_{g,j}^{\text{PluralLLM}})}{K-1} \tag{5.9}$$

$r_{g,j}^{t,\text{Was}} \in [0,1]$, where 0 indicates a perfect distribution match. Lower values indicate better alignment between LLM and group preferences.

Cosine Similarity Reward: Captures directional similarity between preference vectors.

$$r_{g,j}^{t,\text{Cos}} = \frac{y_j^{t,\text{llm}} \cdot p_{g,j}^{\text{PluralLLM}}}{||y_j^{t,\text{llm}}|| \cdot ||p_{g,j}^{\text{PluralLLM}}||} \tag{5.10}$$

$r_{g,j}^{\text{Cos}} \in [-1,1]$, where 1 indicates identical direction, 0 indicates orthogonal, and -1 indicates opposite direction. Higher values indicate better preference alignment.

KL Divergence Reward: Measures information-theoretic alignment.

$$r_{g,j}^{t,\text{KL}} = D_{\text{KL}}\Big(p_{g,j}^{\text{PluralLLM}} \,\Big\|\, y_j^{t,\text{llm}}\Big) = \sum_k p_{g,j,k}^{\text{PluralLLM}} \log \frac{p_{g,j,k}^{\text{PluralLLM}}}{y_{j,k}^{t,\text{llm}}}. \tag{5.11}$$

$r_{g,j}^{\text{KL}} \in [0,\infty)$, where 0 indicates identical distributions, and more positive = greater divergence. Smaller values indicate better alignment.

Ranking-Based Reward Metrics (Preference Ranking Task) These rewards evaluate preference ordering consistency:

Kendall Tau Reward: Measures rank correlation between LLM and PluralLLM orderings.

$$r_{g,j}^{t,\text{Ken}} = \tau\left(\text{rank}\left(y_j^{t,\text{llm}}\right), \text{rank}\left(p_{g,j}^{\text{PluralLLM}}\right)\right). \tag{5.12}$$

$r_{g,j}^{t,\text{Ken}} \in [-1, 1]$, where 1 indicates perfect rank agreement, 0 indicates no correlation, and -1 indicates perfect disagreement. Higher values indicate better ranking alignment.

Borda Reward: Position-weighted scoring based on ranking accuracy.

$$r_{g,j}^{t,\text{Bor}} = \frac{\sum_{k=1}^{K}(K-k+1)\,\mathbb{I}\left[\text{rank}\left(y_j^{t,\text{llm}}\right)_k = \text{rank}\left(p_{g,j}^{\text{PluralLLM}}\right)_k\right]}{K(K+1)/2} \tag{5.13}$$

$r_{g,j}^{t,\text{Bor}} \in [0, 1]$, where 1 indicates perfect position-wise ranking match, and 0 indicates no correct positions. Higher values indicate better ranking quality.

Binary Reward: Simple correctness indicator.

$$r_{g,j}^{t,\text{Bin}} = \mathbb{I}[\text{rank}(y_j^{t,\text{llm}}) = \text{rank}(p_{g,j}^{\text{PluralLLM}})] \tag{5.14}$$

$r_{g,j}^{t,\text{Bin}} \in \{0, 1\}$, where 1 indicates exact ranking match, 0 indicates any disagreement. Binary indicator of perfect alignment.

5.3.4.3 Aggregation Schemes

For each question q_j at FL iteration t, the server aggregates the client-level rewards $\{r_{g_1,j}^t, r_{g_2,j}^t, \ldots, r_{g_l,j}^t\}$ across groups using different strategies:

Average Aggregation:

$$r_{\text{final},j}^t = \frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} r_{g,j}^t \tag{5.15}$$

Provides balanced representation but may mask group-specific needs.

Min Aggregation:

$$r_{\text{final},j}^t = \min_{g \in G_{\text{train}}} r_{g,j}^t \tag{5.16}$$

Ensures no group is left behind but may be overly conservative, limiting overall performance.

Max Aggregation:

$$r^t_{\text{final},j} = \max_{g \in G_{\text{train}}} r^t_{g,j} \tag{5.17}$$

Optimizes for best-case performance but may neglect underrepresented groups.

Adaptive Alpha Aggregation:

$$r^t_{\text{final},j} = \begin{cases} \frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} r^t_{g,j} & \text{if } FI \geq 0.9 \\ \log\left(\frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} \exp(\alpha^t_g \cdot r^t_{g,j})\right) & \text{otherwise} \end{cases} \tag{5.18}$$

Dynamically balances fairness and performance by favoring historically underperforming groups.

The adaptive weights α^t_g are computed using reversed softmax on historical alignment scores:

$$\alpha^t_g = \frac{\exp((1 - h^{t-1}_g)/T)}{\sum_{g' \in G_{\text{train}}} \exp((1 - h^{t-1}_{g'})/T)} \tag{5.19}$$

with temperature $T = 0.1$ and h^{t-1}_g being group g's historical alignment score.

5.3.4.4 Fairness Evaluation Metrics

The Fairness Index (FI) measures reward variation across groups for the same question-response pair:

$$FI = \frac{1}{|X|} \sum_{q_j \in X} \frac{1}{1 + \text{CoV}^2(q_j)} \tag{5.20}$$

where the coefficient of variation for question q_j is:

$$\text{CoV}(q_j) = \frac{\sigma\left(\{r^t_{g,j}\}_{g \in G_{\text{train}}}\right)}{\mu\left(\{r^t_{g,j}\}_{g \in G_{\text{train}}}\right)}. \tag{5.21}$$

$FI \in [0, 1]$, where 1 = perfect fairness (identical rewards across groups), and 0 = maximum

Table 5.1: Fairness evaluation of pluralistic alignment across tasks, rewards, and aggregation strategies. FI denotes the Fairness Index. Alignment scores are reported under multiple metrics; higher is better for all metrics except KL and Was. Both average (Avg AS) and minimum (Min AS) alignment scores are shown. For each column, the best values are highlighted in **bold** (lowest for KL and Was, highest otherwise).

Task	Client Reward	Method	Server Agg.	Fairness Index (FI)						Avg Alignment Score (Avg AS)						Min Alignment Score (Min AS)					
				Was.	Cos.	KL	Ken.	Bor.	Bin.	Was.	Cos.	KL	Ken.	Bor.	Bin.	Was.	Cos.	KL	Ken.	Bor.	Bin.
Preference Prediction Task	—	SFT	—	0.98	0.97	0.88	0.85	0.83	0.97	0.10	0.82	0.4	0.28	0.38	0.23	0.08	0.77	0.55	0.13	0.34	0.23
	WassersteinReward	PPO	Alpha	**0.99**	**0.99**	**0.94**	**0.91**	**0.86**	**1.00**	0.05	0.90	0.26	0.30	0.42	0.22	**0.06**	**0.89**	**0.26**	0.21	0.37	0.22
			Min	0.98	**0.99**	0.93	0.87	0.79	0.90	0.05	**0.91**	**0.22**	0.42	0.44	0.27	**0.06**	**0.89**	0.27	0.34	0.41	**0.28**
			Avg	**0.99**	**0.99**	**0.94**	**0.91**	**0.86**	**1.00**	0.05	0.90	0.26	0.30	0.42	0.22	**0.06**	**0.89**	**0.26**	0.21	0.37	0.22
			Max	**0.99**	**0.99**	0.90	0.88	0.80	0.85	**0.03**	**0.91**	0.23	**0.45**	**0.51**	**0.31**	0.07	**0.89**	0.27	**0.43**	**0.47**	0.31
	CosineReward	PPO	Alpha	**0.99**	**0.99**	0.89	**0.88**	**0.89**	**0.91**	**0.05**	0.92	0.21	0.28	0.42	0.21	**0.06**	0.90	0.27	0.19	0.32	0.19
			Min	**0.99**	**0.99**	0.90	**0.88**	**0.89**	0.80	**0.05**	0.92	0.22	**0.34**	**0.45**	**0.28**	**0.06**	0.90	0.28	**0.21**	**0.34**	**0.22**
			Avg	**0.99**	**0.99**	0.89	**0.88**	**0.89**	**0.91**	**0.05**	0.92	0.21	0.28	0.42	0.21	**0.06**	0.90	0.27	0.19	0.32	0.19
			Max	**0.99**	**0.99**	**0.91**	0.87	0.88	0.88	**0.05**	**0.93**	**0.19**	0.31	0.42	0.22	**0.06**	**0.91**	**0.24**	0.20	**0.34**	0.19
	KLReward	PPO	Alpha	**0.99**	**0.99**	**0.92**	**0.92**	**0.90**	0.78	0.06	**0.92**	0.19	0.40	0.50	0.29	0.07	**0.90**	0.24	0.18	0.38	0.22
			Min	**0.99**	**0.99**	0.91	0.90	0.89	**0.84**	0.06	0.91	**0.17**	**0.43**	**0.51**	**0.33**	0.07	0.89	**0.22**	**0.33**	**0.43**	0.31
			Avg	**0.99**	**0.99**	0.91	0.89	**0.90**	0.76	0.05	0.91	0.19	0.40	0.48	0.27	0.06	0.89	0.26	0.29	0.40	**0.25**
			Max	**0.99**	**0.99**	0.91	0.90	0.86	0.75	**0.04**	0.91	0.19	0.33	0.40	0.20	**0.05**	0.88	0.24	0.22	0.33	0.19
	KendallTauReward	PPO	Alpha	**0.99**	**0.99**	**0.96**	0.90	0.71	**0.91**	0.07	0.75	0.48	0.43	0.38	**0.29**	0.08	0.72	0.55	0.34	0.36	**0.28**
			Min	**0.99**	**0.99**	0.95	0.90	0.75	**0.91**	0.07	0.73	0.49	**0.45**	**0.39**	**0.29**	0.08	0.71	0.54	0.37	0.36	**0.28**
			Avg	**0.99**	**0.99**	0.94	0.90	**0.76**	**0.91**	0.07	0.74	0.48	**0.45**	**0.39**	**0.29**	0.08	0.71	0.54	**0.38**	**0.37**	**0.28**
			Max	**0.99**	**0.99**	0.94	**0.92**	0.73	**0.91**	**0.06**	**0.76**	**0.44**	0.44	0.38	0.28	**0.07**	**0.73**	**0.50**	0.37	0.35	**0.28**
	BordaReward	PPO	Alpha	**0.99**	**0.99**	0.96	**0.89**	**0.71**	0.91	**0.08**	0.73	0.50	0.43	**0.39**	**0.29**	**0.09**	0.69	0.57	0.35	**0.36**	**0.28**
			Min	**0.99**	**0.99**	**0.98**	0.86	0.69	**0.92**	0.09	0.73	0.52	**0.44**	**0.39**	0.28	0.10	**0.70**	0.58	**0.37**	**0.36**	**0.28**
			Avg	**0.99**	**0.99**	0.97	**0.89**	**0.71**	0.91	0.09	0.73	0.51	0.42	0.38	**0.29**	0.10	**0.70**	0.58	0.35	**0.36**	**0.28**
			Max	**0.99**	**0.99**	0.97	**0.89**	**0.71**	0.90	**0.08**	**0.74**	**0.49**	0.44	**0.39**	**0.29**	**0.09**	**0.70**	**0.56**	0.36	**0.36**	**0.28**
	BinaryReward	PPO	Alpha	**0.99**	**0.99**	0.96	**0.89**	**0.68**	**1.00**	**0.08**	**0.74**	**0.52**	**0.42**	**0.38**	0.28	**0.10**	**0.70**	0.60	0.34	0.35	**0.28**
			Min	**0.99**	**0.99**	**0.98**	0.86	0.66	**1.00**	0.10	0.70	0.70	0.37	0.37	0.28	0.11	0.66	0.77	0.30	0.35	**0.28**
			Avg	**0.99**	**0.99**	0.97	**0.89**	0.67	**1.00**	0.09	0.73	0.54	0.42	**0.38**	**0.29**	**0.10**	**0.70**	**0.59**	**0.35**	**0.36**	**0.28**
			Max	**0.99**	**0.99**	0.97	**0.89**	**0.68**	**1.00**	**0.08**	0.73	0.56	0.41	**0.38**	0.28	**0.10**	**0.70**	0.64	0.33	0.34	**0.28**
Preference Ranking Task	—	SFT	—	—	—	—	0.89	0.87	0.83	—	—	—	0.38	0.50	0.31	—	—	—	0.25	0.41	0.27
	KendallTauReward	PPO	Alpha	—	—	—	**0.92**	0.81	0.97	—	—	—	**0.58**	0.47	**0.36**	—	—	—	**0.47**	0.42	**0.31**
			Min	—	—	—	**0.92**	0.81	**0.99**	—	—	—	0.52	0.46	0.30	—	—	—	0.43	0.41	0.28
			Avg	—	—	—	**0.92**	0.82	0.90	—	—	—	0.50	0.48	0.33	—	—	—	0.40	0.40	0.28
			Max	—	—	—	0.91	**0.88**	0.80	—	—	—	0.47	**0.53**	0.35	—	—	—	0.35	**0.44**	0.28
	BordaReward	PPO	Alpha	—	—	—	**0.94**	**0.95**	0.86	—	—	—	**0.53**	**0.61**	**0.39**	—	—	—	0.34	0.45	0.28
			Min	—	—	—	0.92	0.91	**0.89**	—	—	—	0.47	0.53	0.30	—	—	—	**0.36**	0.44	0.28
			Avg	—	—	—	0.93	0.92	0.86	—	—	—	0.49	0.58	**0.39**	—	—	—	0.35	**0.47**	**0.31**
			Max	—	—	—	0.91	0.92	0.78	—	—	—	0.45	0.54	0.32	—	—	—	0.34	0.45	0.28
	BinaryReward	PPO	Alpha	—	—	—	**0.91**	0.89	0.79	—	—	—	**0.49**	0.53	**0.35**	—	—	—	0.33	0.42	0.25
			Min	—	—	—	0.90	0.83	**0.90**	—	—	—	**0.49**	0.49	0.34	—	—	—	**0.39**	0.41	**0.28**
			Avg	—	—	—	**0.91**	0.90	0.79	—	—	—	**0.49**	0.53	**0.35**	—	—	—	0.37	**0.44**	**0.28**
			Max	—	—	—	**0.91**	**0.91**	0.79	—	—	—	0.47	**0.54**	**0.35**	—	—	—	0.35	**0.44**	**0.28**

unfairness. Higher FI values indicate more equitable treatment across demographic groups, while lower values suggest systematic bias favoring certain groups over others. We compute FI separately for each reward metric (Wasserstein, Cosine, KL, Kendall, Borda, Binary) to characterize fairness under different alignment criteria.

5.3.4.5 Preference Probability Prediction Task Results and Analysis

The SFT baseline demonstrates suboptimal performance with fairness indices ranging from 0.83–0.98 and consistently lower alignment scores, highlighting the need for preference-based alignment. Detailed quantitative results are shown in Table 5.1 across multiple reward functions in the Value task.

Reward Function Analysis: Distance-based rewards (Wasserstein, Cosine, KL) substantially outperform ranking-based approaches (Kendall, Borda, Binary). Wasserstein and Cosine rewards with ALPHA aggregation achieve near-optimal fairness ($FI \approx 0.99$) while maintaining strong average alignment scores (Avg AS $\approx$ 0.90–0.95). The minimum alignment scores, crucial for ensuring that no group is left behind, remain competitive (Min AS $\approx$ 0.89–0.94), demonstrating an effective fairness–performance balance.

Dynamic Alpha Aggregation Strategy Impact: Across all distance-based rewards, our adaptive ALPHA aggregation consistently achieves superior fairness indices (often $FI = 0.99$) while preserving competitive average and minimum alignment scores. Compared to static strategies (MIN, AVG, MAX), ALPHA shows two key benefits: (i) it avoids the fairness degradation and worst-group collapse observed with MAX, which can push up average alignment at the cost of marginalized groups; and (ii) it improves worst-group performance relative to AVG, leading to higher Min AS at comparable or better FI. This behavior stems from dynamically reweighting groups based on historical alignment, allowing the server to upweight under-served groups without sacrificing overall utility.

Recommendations: For probability-based preference prediction tasks, we recommend using Wasserstein or Cosine rewards combined with ALPHA aggregation as the default configuration. This combination consistently achieves near-perfect fairness ($FI \approx 0.99$) while maintaining strong average and minimum alignment scores across groups. In settings where heightened sensitivity to distributional mismatch is desired, KL-based rewards with ALPHA aggregation remain competitive but may induce slightly larger fairness variance. Overall, Wasserstein/Cosine + ALPHA provides the most favorable fairness–performance trade-off for modeling calibrated group-level preference probabilities.

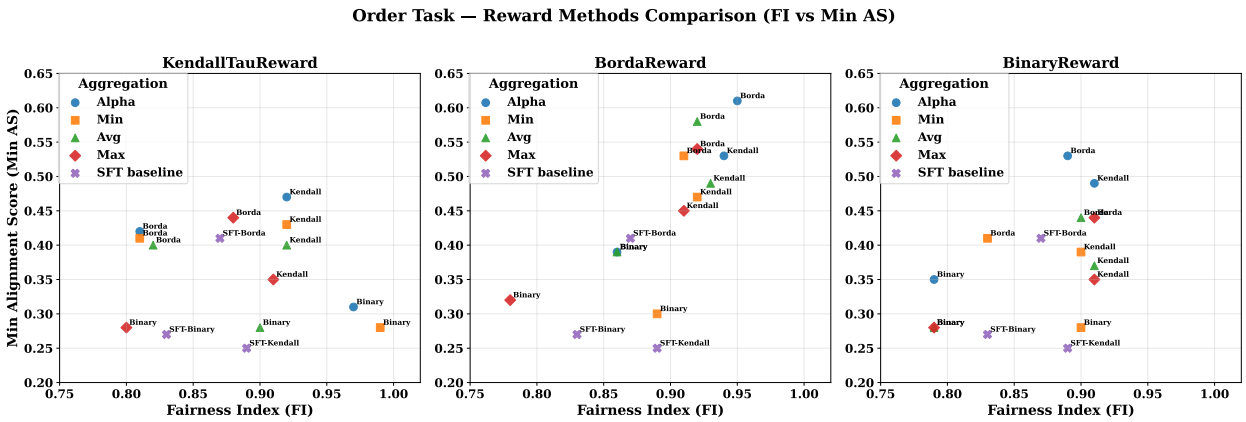


Figure 5.9: **Order task — FI vs. Min AS (worst–group performance) with reward-as-aggregation.** Each subplot fixes *one reward as the main aggregation metric* on the server— *KendallTauReward* (left), *BordaReward* (middle), and *BinaryReward* (right)— and then measures its effect on the three ranking metrics (Kendall, Borda, Binary). Points are server aggregation strategies (ALPHA, MIN, AVG, MAX); SFT baselines are purple crosses. We use the **minimum alignment score (Min AS)** on the y-axis because it reflects the performance of the *worst-served group*, while the x-axis shows the Fairness Index (FI).

5.3.4.6 Preference Ranking Task Results and Analysis

The Order task evaluation focuses on ranking-based metrics (Kendall Tau, Borda, Binary). More details are shown in Table 5.1. The SFT baseline, trained on population-averaged preferences, shows particularly poor performance in ranking tasks with fairness indices of 0.83-0.89 and substantially lower alignment scores (0.31-0.50 average, 0.25-0.41 minimum). This performance degradation in ranking tasks underscores how averaged preference training fails to capture the nuanced ordering preferences that vary significantly across demographic groups.

Ranking Reward Analysis: Alpha aggregation with Kendall Tau rewards achieves the highest fairness index (0.92) and superior average alignment scores (0.58) compared to the SFT baseline (0.38). Borda rewards demonstrate the strongest overall performance, reaching fairness indices up to 0.95 with alpha aggregation and achieving the highest average alignment scores (0.61).

Figure 5.9 demonstrates that **ALPHA aggregation (blue circles) consistently provides the best fairness–performance trade-off,** occupying or approaching the upper-right

quadrant (FI > 0.9, Min AS > 0.3) in all panels. For *KendallTauReward*, ALPHA attains high FI and strong worst-group performance across metrics (Kendall: FI ≈ 0.92, Min AS ≈ 0.47; Borda: $0.81, 0.42$; Binary: $0.97, 0.31$), whereas MIN pushes FI high on Binary (≈ 0.99) but *hurts* Min AS (≈ 0.28). For *BordaReward*, ALPHA achieves the *highest* Min AS across metrics (Kendall ≈ 0.53, Borda ≈ 0.61, Binary ≈ 0.39) with top/tied FI (Kendall ≈ 0.94, Borda ≈ 0.95, Binary ≈ 0.86). For *BinaryReward*, ALPHA again yields the largest Min AS (Kendall ≈ 0.49, Borda ≈ 0.53, Binary ≈ 0.35) with competitive FI, outperforming MIN/AVG/MAX in protecting the worst-served group. Across panels, the SFT baseline underperforms, reinforcing the need for federated preference alignment. Overall, the visualization shows that ALPHA most effectively resolves the fairness–performance tension by *maximizing worst-group (Min AS) performance at high FI.*

Dynamic Alpha Aggregation Strategy Impact: Across all ranking rewards, alpha aggregation maintains competitive performance while consistently achieving better fairness indices than alternative aggregation strategies. Importantly, our evaluation of minimum alignment scores reveals that alpha aggregation successfully prevents the marginalization of lowest-performing groups, maintaining minimum scores (0.31-0.47) that are competitive with or superior to other approaches, while simultaneously achieving higher average performance.

Recommendations: For order-based tasks, we recommend Borda rewards with alpha aggregation, which provides the optimal balance between fairness (0.94-0.95) and alignment performance (0.53-0.61 average). This combination effectively captures group ranking preferences while maintaining equitable treatment across demographic groups.

5.3.4.7 Overall Assessment

Our adaptive alpha aggregation demonstrates superior performance across both task types, consistently achieving the highest fairness indices while maintaining competitive alignment scores. The approach successfully addresses the critical challenge of preventing any group

from being left behind, as evidenced by competitive minimum alignment scores across all evaluation scenarios. These results validate our hypothesis that adaptive weighting based on historical alignment performance provides an effective mechanism for achieving equitable federated learning in preference alignment tasks.

5.3.5 Limitations and Future Work

While our study demonstrates the effectiveness of adaptive alpha aggregation for pluralistic alignment, several areas present natural directions for further research.

Underlying RL Framework. Our current implementation relies on PPO. While effective, PPO can be computationally expensive. Future work should explore more resource-efficient alternatives such as GRPO or DPO, which would allow testing the aggregation strategy across different optimization paradigms and at larger scales.

Model and Dataset Scope. We evaluate on Gemma-2B-it using the Pew Research Global Attitudes dataset, which may be relatively conducive to cross-group alignment. Broader validation on base models and domains with more adversarial or conflicting preferences would provide a stronger stress test of the method's robustness.

Task Generalization. Our experiments focus on multiple-choice Q&A tasks, which offer a controlled setting for evaluation. Extending the framework to diverse tasks such as summarization, dialogue, or code generation would demonstrate its wider applicability and highlight how aggregation impacts more open-ended alignment scenarios.

These considerations do not detract from our main contribution—a systematic evaluation framework with a novel adaptive aggregation scheme—but rather open exciting avenues for expanding its applicability across models, datasets, and tasks.

5.3.6 Conclusion

This evaluation tackles the challenge of aligning LLMs with diverse human preferences in decentralized, federated settings. We showed that *how* group feedback is aggregated is not a minor implementation detail, but a central part of the evaluation protocol that directly governs both fairness and overall alignment performance. Building on PluralLLM's group-specific preference modeling, we proposed an evaluation framework that spans probability prediction and ranking tasks, multiple reward functions, and several aggregation baselines.

Within this framework, our Adaptive Alpha Aggregation dynamically reweights groups based on their historical alignment performance, consistently improving cross-group fairness while maintaining competitive alignment scores. In particular, it raises the performance of the worst-served groups without sacrificing overall utility, providing a practical path toward more pluralistic and equitably aligned LLMs in federated RLHF. This research contributes a valuable evaluation methodology to the field and opens up new avenues for future work, including applying this framework to a broader range of tasks and model architectures.

5.4 Discussion

The two methods proceed in two stages. First, pluralistic alignment can be achieved without the centralization that conventional RLHF assumes. Second, once alignment is federated, how group preferences are *combined* determines fairness. PluralLLM shows that group preferences can be learned collaboratively without pooling raw data, and that a lightweight federated preference predictor is a viable, efficient stand-in for the heavy reward-modeling machinery of RLHF, keeping preference data with the groups that own it while still producing a model that generalizes to communities it has not seen. The aggregation study then shows that federation also affects fairness. Given decentralized per-group rewards, static rules encode a value judgment: averaging can mask a marginalized group, and maximizing can sacrifice the worst-served group to raise the mean, whereas an adaptive rule that attends

to which groups are currently underserved can lift the worst-off without giving up overall alignment quality. The two papers together give a path to alignment that is both privacy-preserving and fairness-aware, and an evaluation protocol under which whose preferences are honored is measured, not assumed.

Fairness here is again the equity of a distribution, now over the *values* a model represents, and the quantity protected is the alignment of the worst-served group, the same policy of attending to whoever is falling behind that FAIRO applied to human experience. Privacy is again preserved by keeping sensitive data decentralized and learning over it in a federated loop, as in Chapter 3. The object being aligned is the same human preference that Chapter 4 modeled for the individual, now modeled for the group. Chapter 6 draws the full argument together and looks ahead.

Chapter 6

Conclusion

6.1 Summary

Every system studied in this dissertation senses human state, infers human preference, and adapts in response: smart environments, cyber-physical systems, and conversational large language models. Systems of this kind allocate something among the people in the loop: comfort in a shared room, exposure to an inference attack, representation in a model's outputs. What they allocate is rarely measured. This work treats human experience, privacy risk, and preference as objectives with metrics of their own, not as constraints attached to a system whose real goal is task performance.

The chapters follow this order. Chapter 2 began with *fairness in decision-making*, formalizing fair treatment across a group of people as the goal of minimizing the discrepancy in their accumulated experience over time, and pursuing that goal from two directions: FinA, an optimization view grounded in loss aversion, and FAIRO, a learning view grounded in the temporal abstraction of the Options framework. Chapter 3 applied the same equity view to *privacy*, treating privacy risk in federated learning as a distribution of risk, and enforcing

fairness-in-privacy through FinP's Hessian-ranked, Lipschitz-based client regularization and risk-aware server aggregation. Chapter 4 moved from the group to the individual, showing with PACIFIC that stable Big-Five personality traits are a more reliable basis for *personalization* than raw interaction history, and that the bottleneck lies in constructing the right context, not in the model's capacity to reason about it. Chapter 5 returned to the collective, studying *pluralistic alignment* of language models through PluralLLM's federated preference learning and a systematic evaluation showing that how group preferences are aggregated determines fairness across communities.

6.2 Synthesis of Contributions

Beyond the specific results of each chapter, three ideas recur across the four settings and constitute the central contribution of this dissertation.

Many human-centered failures are problems of *distribution*, not only of magnitude. Fairness in decision-making is the spread of experience across people; fairness in privacy is the spread of leakage risk; fairness in alignment is the spread of representation across groups. In each case it matters how good an outcome is on average, how it is distributed, and who bears the worst of it. Protecting the worst-off, not only the mean, is what makes a system fair. FAIRO's policy of repeatedly attending to whoever is falling behind, FinP's harder regularization of the clients most exposed, and adaptive aggregation's upweighting of the worst-served group all follow this view.

These gains need not be paid for in utility. Across all four settings, treating human experience, privacy risk, and preference as explicit objectives improved them without a corresponding loss in task performance: fairness was achieved without sacrificing thermal comfort or resource sufficiency, privacy equity was achieved within a marginal utility cost and even reduced overall leakage, and pluralistic alignment was achieved while preserving alignment

quality. With the right formulation, the trade-off between serving people and serving the task is weaker than it is often assumed to be.

Adapting faithfully to people requires a useful *abstraction* of them. PACIFIC's central lesson is that a compact, stable trait profile serves an individual better than a lossless replay of their history. The same kind of compression appears in every chapter: FinP reads the geometry of the loss surface, not the raw data, to locate risk, and PluralLLM learns a lightweight predictor of a group's preferences without centralizing them. In each case the useful representation of a person or group is a compressed, purpose-built abstraction, not the raw record.

6.3 Limitations and Future Directions

Several directions follow directly from the limitations of the frameworks presented here.

Unifying formal guarantees with empirical equity. FinP establishes fairness-in-privacy empirically, while differential privacy offers formal guarantees but is blind to per-client geometry. Combining the two, using geometry-aware, per-client privacy budgets that allocate protection where the risk actually concentrates, would yield privacy that is both provable and equitable.

Auditing and correcting persona-level bias. PACIFIC surfaced a systematic bias in which "low" expressions of personality traits are recovered far less reliably than their socially desirable opposites, so that for a large part of the trait spectrum personalization defaults to the normatively approved answer. Diagnosing where this bias enters the alignment pipeline, and correcting it so that personalization does not overwrite the users it is meant to serve, is important future work, and it ties personalization back to fairness.

Broadening pluralistic alignment. The alignment work is demonstrated on multiple-choice question-answering with a fixed reinforcement-learning backbone and a survey-based

preference dataset. Extending it to open-ended tasks such as summarization, dialogue, and code generation; to more resource-efficient optimization such as GRPO or DPO; and to domains with more genuinely adversarial or conflicting preferences would test both federated preference learning and adaptive aggregation, and would move federated RLHF from a controlled evaluation toward practical deployment.

Toward a unified human-in-the-loop framework. Each chapter of this dissertation treats one property (fairness, privacy, personalization, or alignment) as the primary objective while holding the others fixed. A joint treatment remains open: a single human-in-the-loop system that distributes experience fairly, distributes privacy risk equitably, personalizes faithfully to each individual, and aligns to the plural values of the groups it serves, trading these objectives off explicitly instead of one at a time. Building the formulations and evaluation protocols for that joint problem continues this work.

6.4 Closing Remarks

A system that senses, infers, and adapts to people is responsible for how it treats them, for what it discloses about them, and for how accurately it represents them. Across four settings, and with results to support it, this dissertation has argued that those responsibilities can be written into the objective and measured directly, and that meeting them does not require giving up the performance that made the systems useful in the first place.

Bibliography

General data protection regulation. `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation`, 2018.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

J Stacy Adams. Towards an understanding of inequity. *The journal of abnormal and social psychology*, 67(5):422, 1963.

Anuradha Annaswamy, Karl Johansson, and George Pappas. Control for societal-scale challenges roadmap 2030, 2023a.

Anuradha M. Annaswamy, Pramod P. Khargonekar, Françoise Lamnabhi-Lagarrigue, and Sarah K. Spurgeon. *Cyber-Physical-Human Systems: Fundamentals and Applications*. John Wiley & Sons, Inc., 2023b.

ASHRAE/ANSI Standard 55-2010 American Society of Heating, Refrigerating, and Air-Conditioning Engineers. Thermal environmental conditions for human occupancy. *Inc.Atlanta, GA, USA*, 2010.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Adda-Akram Bendoukha, Didem Demirag, Nesrine Kaaniche, Aymen Boudguiga, Renaud Sirdey, and Sébastien Gambs. Towards privacy-preserving and fairness-aware federated learning framework. In *Privacy Enhancing Technologies (PETs)*, volume 2025, pages 845–865, 2025.

Big Five Personality Test. Big Five Personality Test. https://openpsychometrics.org/tests/IPIP-BFFM/, August 2019.

Stephen R Briggs. Assessing the five-factor model of personality description. *Journal of personality*, 60(2):253–293, 1992.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Iván Cantador, Ignacio Fernández-Tobías, Alejandro Bellogín, Michal Kosinski, and David Stillwell. Relating personality types with user preferences in multiple entertainment domains. In *UMAP Workshops*, volume 997, 2013.

Shuirong Cao, Ruoxi Cheng, and Zhiqiang Wang. Agr: Age group fairness reward for bias mitigation in llms. *arXiv preprint arXiv:2409.04340*, 2024.

Robert G. Carroll. Pulmonary system. In *Elsevier's Integrated Physiology*, chapter 10, pages 99–115. Elsevier, 2007.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Hongyan Chang, Brandon Edwards, Anindya S Paul, and Reza Shokri. Efficient privacy auditing in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 307–323, 2024.

Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. pfl-bench: A comprehensive benchmark for personalized federated learning. *Advances in Neural Information Processing Systems*, 35:9344–9360, 2022.

Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. Proxona: Supporting creators' sensemaking and ideation with llm-powered audience personas. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3714034. URL https://doi.org/10.1145/3706598.3714034.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Robert Cialdini. *Influence: The Psychology of Persuasion*. Harper Collins, 2007.

Robert Cialdini. *Presuasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.

Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–308, 2020.

Federico Concone, Cedric Ferdico, Giuseppe Lo Re, and Marco Morana. A federated learning approach for distributed human activity recognition. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 269–274. IEEE, 2022.

Karen S Cook and Richard Marc Emerson. Social exchange theory. 1987.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pages 2185–2195. PMLR, 2020.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

Mirosława Czerniawska and Joanna Szydło. Do values relate to personality traits and if so, in what way?–analysis of relationships. *Psychology research and behavior management*, pages 511–527, 2021.

Boele De Raad. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers, 2000.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. Federated learning based on dynamic regularization. In *International conference on learning representations*, 2021.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Salma Elmalaki. Fair-iot: Fairness-aware human-in-the-loop reinforcement learning for harnessing human variability in personalized iot. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 119–132, 2021.

Salma Elmalaki, Huey-Ru Tsai, and Mani Srivastava. Sentio: Driver-in-the-loop forward collision warning using multisample reinforcement learning. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 28–40, 2018.

Salma Elmalaki, Bo-Jhang Ho, Moustafa Alzantot, Yasser Shoukry, and Mani Srivastava. Vindico: Privacy safeguard against adaptation based spyware in human-in-the-loop iot. *arXiv preprint arXiv:2202.01348*, 2022.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7494–7502, 2023.

Poul O Fanger. Thermal comfort. analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

Grace Fox, Trevor Clohessy, Lisa van der Werff, Pierangelo Rosati, and Theo Lynn. Exploring the competing influences of privacy concerns and positive beliefs on citizen acceptance of contact tracing mobile applications. *Computers in Human Behavior*, 121:106806, 2021.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

Michael Gerber. energyplus energy simulation software. 2014.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in neural information processing systems*, pages 2600–2609, 2018.

Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 116–116, 2018.

Lewis R. Goldberg. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42, 1992. doi: 10.1037/1040-3590.4.1.26.

Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge, 2013.

Avi Goldfarb and Catherine Tucker. Shifts in privacy concerns. *American Economic Review*, 102(3):349–53, 2012.

Yuhao Gu, Yuebin Bai, and Shubin Xu. Cs-mia: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications*, 67:103201, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581, 2012.

Florence Ho, Rúben Geraldes, Artur Gonçalves, Bastien Rigault, Benjamin Sportich, Daisuke Kubo, Marc Cavazza, and Helmut Prendinger. Decentralized multi-agent path finding for uav traffic management. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):997–1008, 2020.

George C Homans. Social behavior: Its elementary forms. 1974.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1102–1107. IEEE, 2021.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Hongsheng Hu, Xuyun Zhang, Zoran Salcic, Lichao Sun, Kim-Kwang Raymond Choo, and Gillian Dobbie. Source inference attacks: Beyond membership inference attacks in federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2023.

Rong Hu and Pearl Pu. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 197–204, 2011.

SHI Huaizhou, R Venkatesha Prasad, Ertan Onur, and IGMM Niemegeers. Fairness in wireless networks: Issues, measures and challenges. *IEEE Communications Surveys & Tutorials*, 16(1):5–24, 2013.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336, 2018.

Nicole MA Huijts, Eric JE Molin, and Linda Steg. Psychological factors influencing sustainable energy technology acceptance: A review-based comprehensive framework. *Renewable and sustainable energy reviews*, 16(1):525–531, 2012.

Alex Iosup, Xiaoyun Zhu, Arif Merchant, Eva Kalyvianaki, Martina Maggio, Simon Spinner, Tarek Abdelzaher, Ole Mengshoel, and Sara Bouchenak. Self-awareness of cloud applications. *Self-Aware Computing Systems*, pages 575–610, 2017.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pages 1617–1626, 2017.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. A quantitative measure of fairness and discrimination. *ACM Transaction on Computer System*, 1984.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

Jiechuan Jiang and Zongqing Lu. Learning fairness in multi-agent systems. In *Advances in Neural Information Processing Systems*, pages 13854–13865, 2019.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJgIPJBFvH`.

Haiming Jin, Lu Su, Bolin Ding, Klara Nahrstedt, and Nikita Borisov. Enabling privacy-preserving incentives for mobile crowd sensing systems. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, pages 344–353. IEEE, 2016.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

Wooyoung Jung and Farrokh Jazizadeh. Towards integration of doppler radar sensors into personalized thermoregulation-based control of hvac. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, page 21. ACM, 2017.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, pages 2227–2236, 2018.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://www.aclweb.org/anthology/2020.emnlp-main.550`.

Pramod P Khargonekar and Meera Sampath. A framework for ethics in cyber-physical-human systems. *IFAC-PapersOnLine*, 53(2):17008–17015, 2020.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5260–5271, 2024.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies*, 2022.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management

that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3365–3376, 2017.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.999. URL https://aclanthology.org/2025.acl-long.999/.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

MATLAB. Water distribution system scheduling using reinforcement learning. https://www.mathworks.com/help/reinforcement-learning/ug/water-distribution-scheduling-system.html, March 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research*, 9(5):583–597, 2002.

Deirdre K Mulligan, Colin Koopman, and Nick Doty. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160118, 2016.

Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3):1–37, 2022.

Huansheng Ning, Sahraoui Dhelim, and Nyothiri Aung. Personet: Friend recommendation system based on big-five personality traits and hybrid filtering. *IEEE Transactions on Computational Social Systems*, 6(3):394–402, 2019.

Northpointe. Practitioner's guide to COMPAS core. `https://www.equivant.com/practitioners-guide-to-compas-core/`, 2015.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Philadelphia Government. Gallons used per person per day. `https://water.phila.gov/pool/files/home-water-use-ig5.pdf`, 2023.

Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. Improving fairness in ai models on electronic health records: The case for federated learning methods. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1599–1608, 2023.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*, volume 110, pages 91–95, 2020.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

Lillian J Ratliff and Tanner Fiez. Adaptive incentive design. *IEEE Transactions on Automatic Control*, 66(8):3871–3878, 2020.

Lillian J Ratliff, Roy Dong, Shreyas Sekar, and Tanner Fiez. A perspective on incentive design: Challenges and opportunities. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):1–34, 2018.

Mark Redmond. Social exchange theory. 2015.

Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

Peter J Rentfrow, Lewis R Goldberg, and Ran Zilca. Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of personality*, 79(2): 223–258, 2011.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C54S4K.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Steve Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly. Zero trust architecture. Technical Report SP 800-207, National Institute of Standards and Technology (NIST), Gaithersburg, MD, 2020. URL `https://doi.org/10.6028/NIST.SP.800-207`.

Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533, 2024.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, 2024.

John Schulman et al. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Giovanni M Sechi, Riccardo Zucca, and Paola Zuddas. Water costs allocation in complex systems using a cooperative game theory approach. *Water Resources Management*, 27: 1781–1796, 2013.

Khotso Selialia, Yasra Chandio, and Fatima M Anwar. Mitigating group bias in federated learning for heterogeneous devices. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1043–1054, 2024.

Amartya Sen. *On economic inequality*. Oxford university press, 1997.

Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Eun-Jeong Shin, Roberto Yus, Sharad Mehrotra, and Nalini Venkatasubramanian. Exploring fairness in participatory thermal comfort control in smart buildings. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, pages 1–10, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

Herbert Alexander Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1997.

M Joseph Sirgy. Using self-congruity and ideal congruity to predict purchase motivation. *Journal of business Research*, 13(3):195–206, 1985.

M Joseph Sirgy. Self-congruity theory in consumer behavior: A little history. *Journal of Global Scholars of Marketing Science*, 28(2):197–207, 2018.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. Pluralllm: pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pages 64–69, 2025a.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. A systematic evaluation of preference aggregation in federated rlhf for pluralistic alignment of llms. *arXiv preprint arXiv:2512.08786*, 2025b. NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. Appa: Adaptive preference pluralistic alignment for fair federated rlhf of llms. *arXiv preprint arXiv:2604.04261*, 2026.

Spectrum News Staff. Massive data breach exposes millions of americans. `https://spectrumlocalnews.com/nys/central-ny/news/2024/09/04/data-breach-exposes-americans-information`, September 2024.

Charles Stangor. Conflict, Cooperation, Morality, and Fairness. In *Principles of Social Psychology - 1st International Edition*. BCcampus Pressbooks, 2014.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112 (1-2):181–211, 1999.

Janos Sztipanovits, Xenofon Koutsoukos, Gabor Karsai, Shankar Sastry, Claire Tomlin, Werner Damm, Martin Fränzle, Jochem Rieger, Alexander Pretschner, and Frank Köster. Science of design for societal-scale cyber-physical systems: challenges and opportunities. *Cyber-Physical Systems*, 5(3):145–172, 2019.

Mojtaba Taherisadr and Salma Elmalaki. Pearl: Personalized privacy of human-centric systems using early-exit reinforcement learning. *arXiv preprint arXiv:2403.05864*, 2024.

Mojtaba Taherisadr, Stelios Andrew Stavroulakis, and Salma Elmalaki. Adaparl: Adaptive privacy-aware reinforcement learning for sequential decision making human-in-the-loop systems. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, pages 262–274. ACM, 2023.

Maya Tamir. Don't worry, be happy? neuroticism, trait-consistent affect regulation, and performance. *Journal of personality and social psychology*, 89(3):449, 2005.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Shogo Terai, Shizuka Shirai, Mehrasa Alizadeh, Ryosuke Kawamura, Noriko Takemura, Yuki Uranishi, Haruo Takemura, and Hajime Nagahara. Detecting learner drowsiness based on facial expressions and head movements in online courses. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pages 124–125, 2020.

John W Thibaut and Harold Harding Kelley. *The social psychology of groups*. Routledge, 1959.

Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 15–28, 2021.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Olivier Polleux, Kashif Rasul, Lysandre Debut, and Omar Sanseviero. TRL: Transformer Reinforcement Learning. `https://github.com/huggingface/trl`, 2022.

Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, 2022.

Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3345–3355, 2024.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

Zhenhuan Yang, Yingqiang Ge, Congzhe Su, Dingxian Wang, Xiaoting Zhao, and Yiming Ying. Fairness-aware differentially private collaborative filtering. In *Companion Proceedings of the ACM Web Conference 2023*, pages 927–931, 2023.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li, et al. Too good to be bad: On the failure of llms to role-play villains. *arXiv preprint arXiv:2511.04962*, 2025.

Yiding Yu, Taotao Wang, and Soung Chang Liew. Deep-reinforcement learning multiple access for heterogeneous wireless networks. *IEEE Journal on Selected Areas in Communications*, 37(6):1277–1290, 2019.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023a.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*, 2025.

Tianyu Zhao and Salma Elmalaki. Fina: Fairness of adverse effects in decision-making of human-cyber-physical-system. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*, pages 202–211. IEEE, 2024.

Tianyu Zhao, Mojtaba Taherisadr, and Salma Elmalaki. Fairo: Fairness-aware sequential decision making for human-in-the-loop cps. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*, pages 87–98. IEEE, 2024.

Tianyu Zhao, Siqi Li, Yasser Shoukry, and Salma Elmalaki. Pacific: Can llms discern the traits influencing your preferences? evaluating personality-driven preference alignment in llms. *arXiv preprint arXiv:2602.07181*, 2026a.

Tianyu Zhao, Mahmoud Srewa, and Salma Elmalaki. FinP: Fairness-in-privacy in federated learning by addressing disparities in privacy risk. *Proceedings on Privacy Enhancing Technologies*, 2026(4):587–607, 2026b. doi: 10.56553/popets-2026-0136.

Yiqi Zhao, Ziyan An, Xuqing Gao, Ayan Mukhopadhyay, and Meiyi Ma. Fairguard: Harness logic-based fairness rules in smart cities. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, pages 105–116, 2023b.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

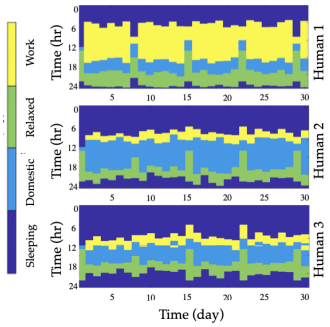
Gongxi Zhu, Donghao Li, Hanlin Gu, Yuxing Han, Yuan Yao, Lixin Fan, and Qiang Yang. Evaluating membership inference attacks and defenses in federated learning. *arXiv preprint arXiv:2402.06289*, 2024.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

Appendix A

FAIRO Appendix

Appendix



Histogram	Rooms	FAIRO	Average	RR	1 DQN
Satisfaction Gaussian fitting μ,σ^2	Room 1	0.642, 0.047	0.452, 0.030	0.748, 0.021	0.657, 0.142
	Room 2	0.649, 0.050	0.506, 0.027	0.785, 0.023	0.641, 0.084
	Room 3	0.641, 0.047	0.487, 0.030	0.698, 0.024	0.680, 0.152
Satisfaction JSD	Room 1&2	0.022 bits	0.399 bits	0.269 bits	0.115 bits
	Room 1&3	**0.002** bits	0.175 bits	0.443 bits	0.020 bits
	Room 2&3	**0.014** bits	0.080 bits	0.835 bits	0.121 bits
	avg. JSD	**0.013** bits	0.218 bits	0.516 bits	0.085 bits
PMV Gaussian fitting μ,σ^2	Room 1	0.346, 0.490	0.369, 0.430	0.367, 0.590	0.336, 0.502
	Room 2	-0.014, 0.612	0.021, 0.498	0.021, 0.701	-0.030, 0.642
	Room 3	-0.099, 0.673	-0.065, 0.564	-0.052, 0.732	-0.120, 0.723
PMV JSD	Room 1&2	0.088 bits	0.112 bits	**0.071** bits	0.089 bits
	Room 1&3	0.119 bits	0.152 bits	**0.099** bits	0.119 bits
	Room 2&3	0.012 bits	0.023 bits	**0.014** bits	0.014 bits
	avg. JSD	0.073 bits	0.096 bits	**0.063** bits	0.074 bits

Figure A.1: Left: Human activities, Right: Comparison of the satisfaction values $\frac{v}{u+v}$ and the PMV across all the approaches in application 1. Compared with all methods, the JSD of FAIRO satisfaction is reduced by 94.0% from Average Approach, reduced by 97.5% from RR, and reduced by 84.7% from 1 DQN. Similarly, FAIRO's PMV JSD is reduced by 24.0% from Average Approach, increased by 15.9% from RR, and reduced by 1.4% from 1 DQN.

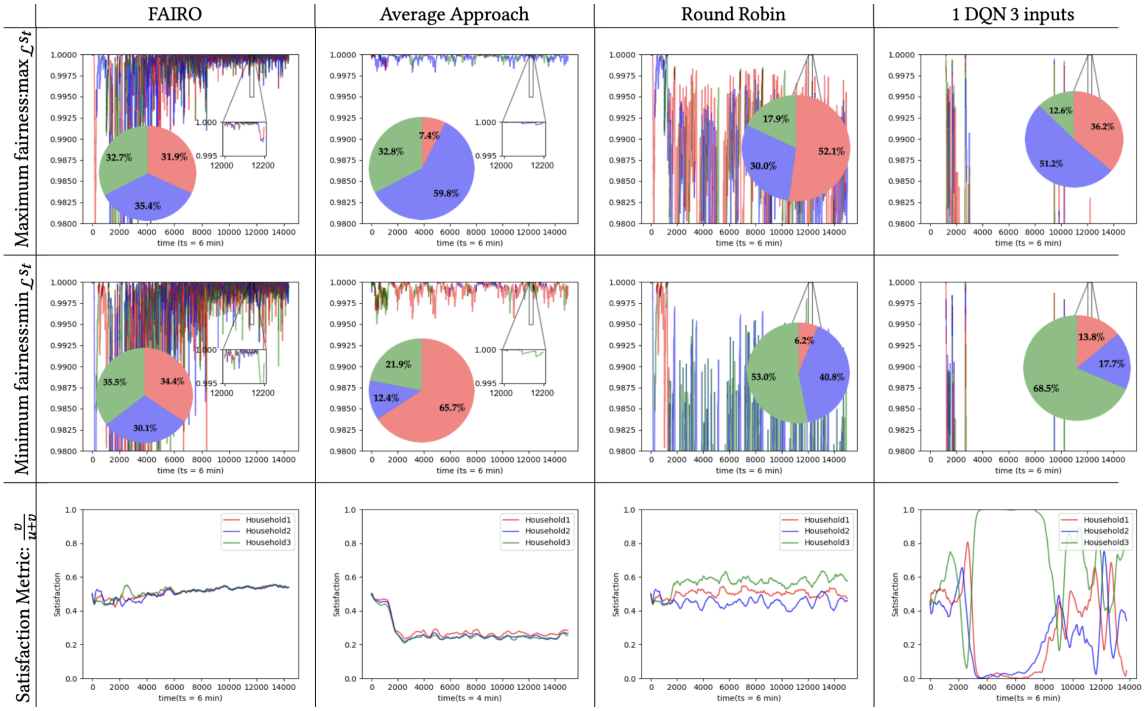


Figure A.2: Fairness state and satisfaction across all approaches for application 2. First and second rows show the maximum and minimum value of s_t. The fairness states results with $15k$ samples equivalent to 62.5 simulated days. Using FAIRO, the average absolute difference between the probabilities of *equalized odds* across the 3 households is reduced by 31.9%, 31.0%, 37.1%, 27.6%, and 32.9% from Weighted Average, Weighted RR, Average, RR,and 1 DQN 3 inputs respectively with an average of 32.1. Third row reports the satisfaction values across the three households. Satisfaction values in FAIRO are closer compared with other approaches with satisfaction values around 0.55. RR satisfaction values are from 0.6 to 0.4 with different between households. The Weighted Average Approach satisfaction values are close, but values stay below 0.3. 1 DQN-3 inputs has unstable fluctuations across 3 households.

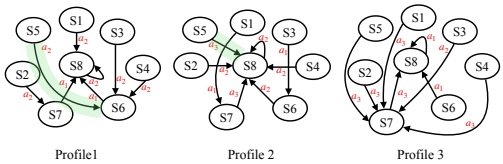
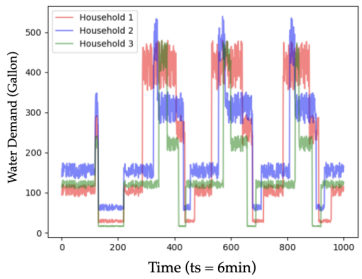


Figure A.3: MDP for three human profiles in VR learning environment.



Histogram	Rooms	FAIRO	Weighted Average	Weighted RR	Average	RR	1 DQN
Satisfaction Gaussian fitting μ,σ^2	Room 1	**0.540**, 0.037	0.267, 0.021	0.459, 0.014	0.214, 0.100	0.496, 0.061	0.342, 0.280
	Room 2	**0.540**, 0.029	0.249, 0.022	0.445, 0.014	0.205, 0.085	0.462, 0.050	0.315, 0.257
	Room 3	0.535, 0.048	0.244, 0.020	0.429, 0.013	0.495, 0.105	0.596, 0.052	**0.636**, 0.241
Satisfaction JSD	Room 1&2	**0.016** bits	0.091 bits	0.079 bits	0.088 bits	0.094 bits	0.048 bits
	Room 1&3	**0.011** bits	0.150 bits	0.427 bits	0.654 bits	0.390 bits	0.237 bits
	Room 2&3	0.024 bits	**0.010** bits	0.165 bits	0.756 bits	0.644 bits	0.288 bits
	JS Average	**0.017** bits	0.084 bits	0.224 bits	0.499 bits	0.376 bits	0.191 bits
Balance Rate (BR)	avg.>80%BR (# samples > 80%BR)	**0.535(1606)**	0.254(763)	0.446(1336)	0.305(915)	0.517(1552)	0.432(1296)

Figure A.4: Left: Water demand profile for three households during three days, Right: Comparison of the satisfaction values $\frac{v}{u+v}$ and the balance rate (BR) across all the approaches in application 1. In particular, for satisfaction, compared with other methods, FAIRO JSD is reduced by 79.8%, 92.4%, 96.5%, 95.5%, and 91.1% from Weighted Average Approach, Weighted RR, Average Approach, RR, and 1 DQN 3 inputs methods respectively with an average of 91.06%. In terms of BR, FAIRO has the highest average $> 80\%$ BR value (0.535). Compared with other methods, FAIRO's average $> 80\%$ BR is improved by 110.6%, 20.0%, 75.4%, 3.5%, and 23.8% from Weighted Average Approach, Weighted RR, Average Approach, RR, and 1 DQN 3 inputs methods respectively with an average of 46.66%.

Histogram	Rooms	FAIRO	Average	RR	1 DQN
Satisfaction Gaussian fitting μ, σ^2	Room 1	0.639, 0.017	0.761, 0.033	0.594, 0.015	0.692, 0.062
	Room 2	0.647, 0.020	0.821, 0.020	0.671, 0.018	0.787, 0.050
	Room 3	0.638, 0.018	0.554, 0.020	0.583, 0.012	0.604, 0.094
Satisfaction JSD	Room 1&2	**0.022** bits	0.528 bits	0.801 bits	0.396 bits
	Room 1&3	**0.000** bits	1 bits	0.060 bits	0.229 bits
	Room 2&3	**0.021** bits	1 bits	0.884 bits	0.657 bits
	JS Average	**0.094** bits	0.843 bits	0.582 bits	0.427 bits
Learning Experience (LE)	Overlap Samples %	0.488	0.438	0.507	**0.497**

Table A.1: Comparison of the satisfaction values $\frac{v}{u+v}$ and the learning experience (LE) in application 3. FAIRO satisfaction JSD is reduced by 88.8%, 83.8%, and 78.0% from Average Approach, RR, and 1 DQN respectively.

Figure A.5: Fairness state and satisfaction metric across all different approaches for application 3. First and second rows show the maximum and minimum value of s_t. FAIRO achieves *equal opportunity* probabilities 42.1%, 24.0%, and 33.9% for human #1, #2, and #3 respectively. As for *equalized odds* probabilities, FAIRO reports 24.3%, 49.4%, and 26.3% for humans #1, #2, and #3 respectively. While the difference in these values across the three humans is not as small as we had in the other two applications, we observe that this difference in FAIRO is better than the other approaches.

Appendix B

F*in*P Appendix

B.1 Datasets and Setup

We evaluate F*in*P using the UCI HAR[Reyes-Ortiz et al., 2013] and CIFAR-10 datasets [Krizhevsky et al., 2009] utilizing federated learning with non-IID data partitions and standard model architectures (TCN for HAR, CNN and ResNet56 for CIFAR-10).

Setup for Human Activity Recognition We utilized the UCI HAR Dataset [Reyes-Ortiz et al., 2013], a widely used dataset in activity recognition research, especially in FL [Concone et al., 2022, Tu et al., 2021]. The dataset includes sensor data from 30 subjects (aged 19–48) performing six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. The data was collected using a Samsung Galaxy S II smartphone worn on the waist, capturing readings from both the accelerometer and gyroscope sensors. Each subject in the dataset was treated as an individual client in the FL setup, preserving the data's unique activity patterns and non-IID nature. We allocated 70% of each client's data for training using 5-fold cross-validation and 30% for testing, enabling evaluation of the model on independently collected test data. Data preprocessing involved applying noise filters to the raw signals and segmenting the data using a sliding window approach with a window length of 2.56 seconds and a 50% overlap, resulting in 128 readings per window. We selected the HAR dataset for evaluation F*in*P due to its inherited non-IID structure.

We trained the model in a federated learning setting using the Federated Averaging Algorithm (FedAvg) aggregation method over 20 global communication rounds. Each client trained locally with a batch size of 64, a learning rate of 0.001 using Adam optimizer, 1 local epochs per round, and an impact factor β of 2. These parameters ensured balanced model updates from each client while maintaining computational efficiency across the federated network. Each local model (one per subject) analyzes its time-series sensor data using Temporal Convolutional Network (TCN) modelBai et al. [2018]. The TCN model, designed for time-series data, uses causal convolutions to capture temporal dependencies while preserving

sequence order. The architecture includes two convolutional layers, each followed by max-pooling, with a final fully connected layer for classifying the six activity classes.

Setup for CIFAR-10 The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. We use the Dirichlet distribution $Dir(\alpha)$ to divide the CIFAR-10 dataset into K unbalanced subsets similar to previous work in the literature [Mendieta et al., 2022, Hu et al., 2021], with $\alpha = 0.5$. Figure 3.2 demonstrates how the data are distributed among clients with $\alpha = 0.1$. We created 10 clients and employed ResNet56 [He et al., 2016] as the local model. Similar to the setup in HAR, we trained the model over 20 global communication rounds. Each client is trained locally with an impact factor β of 0.1 as described in Equation 3.5. As for ablation experiment in β, we evaluated multiple values of $\beta = \{0.05, 0.1, 0.3, 0.5\}$ described in Equation 3.5. We used the same CNN proposed in Hu et al. [2023] as the local model.

Setup for FEMNIST We evaluate on FEMNIST from Hugging Face (flwrlabs/femnist), a handwritten character recognition benchmark with 62 classes and natural non-IID structure induced by writer identity. We simulate K federated clients by sampling K distinct writers without replacement; each client owns only that writer's examples. Images are converted to single-channel tensors and normalized with MNIST statistics (μ=0.1307, σ=0.3081). For each client, we perform an 80/20 train–test split at the writer level, yielding client-local heterogeneity in both label and feature distributions.

We employ a lightweight CNN, FEMNISTNet, tailored to 28×28 grayscale inputs. The architecture comprises two convolutional blocks (32 and 64 channels, 3×3 kernels, GroupNorm, GELU), each followed by learnable strided downsampling (28→14→7), and two fully connected layers (3136→256→62). and local training uses SGD with learning rate 0.02. The model has on the order of ∼3M trainable parameters and outputs unnormalized class logits

for 62-way classification with cross-entropy loss. This design balances capacity and efficiency for per-client federated optimization and curvature-based collaboration metrics.

B.2 Experiment Results on HAR

B.2.1 HAR Results with PCA as server aggregation

Figures B.1 and B.2 show the average SIA accuracy and the equal opportunity difference (EOD) per client using HAR dataset with PCA-based server aggregation, respectively.

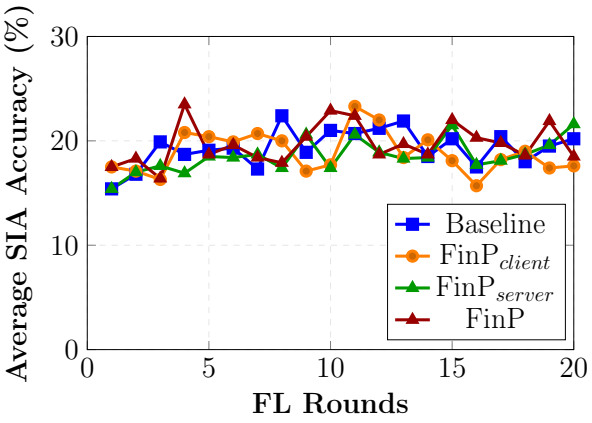


Figure B.1: Average SIA accuracy of Baseline, FinP $_{\text{client}}$, FinP $_{\text{server}}$, and FinP with PCA on HAR dataset.

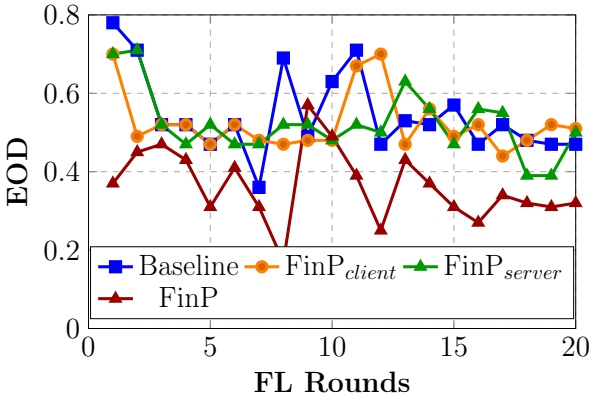
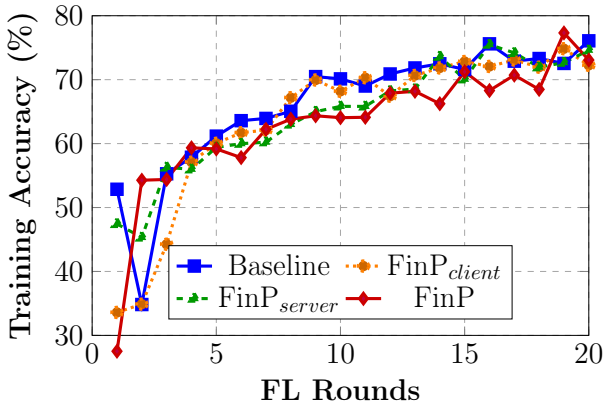


Figure B.2: Equal opportunity difference (EOD) on HAR dataset and FinP with PCA.

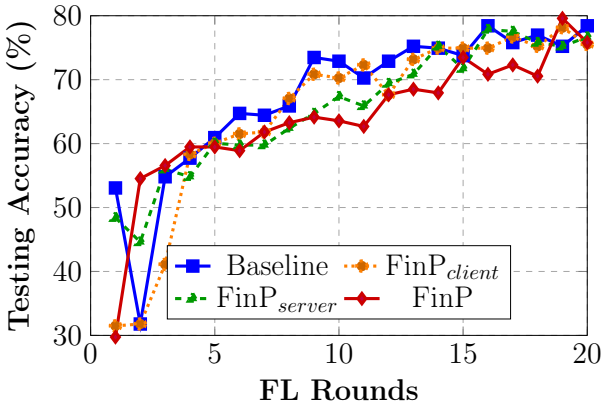
B.2.2 HAR with Adaptive Lightweight Aggregation (ALA)

Figure B.3 shows the training and testing accuracy for HAR dataset using the lightweight server aggregation.

Figures B.4 and B.5 describe the FI(SIA) and FI(loss) for HAR dataset using ALA aggregation. Figure B.6 shows the average SIA accuracy in HAR with ALA while Figure B.7 highlights the Equal Opportunity Difference (EOD) in HAR with ALA.



(a) Training accuracy.



(b) Testing accuracy.

Figure B.3: Global model classification accuracy using HAR with ALA.

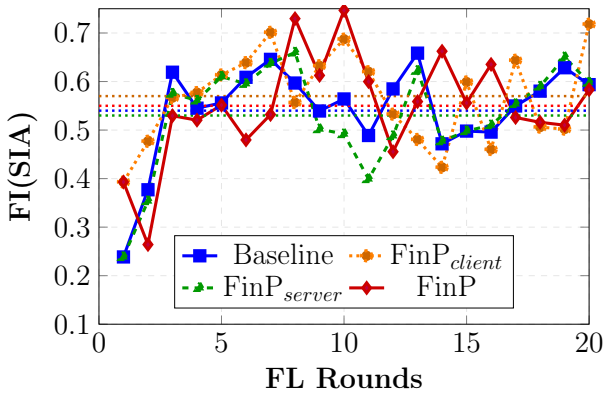


Figure B.4: Disparity of SIA accuracy (FI(SIA)) among clients using HAR dataset with ALA.

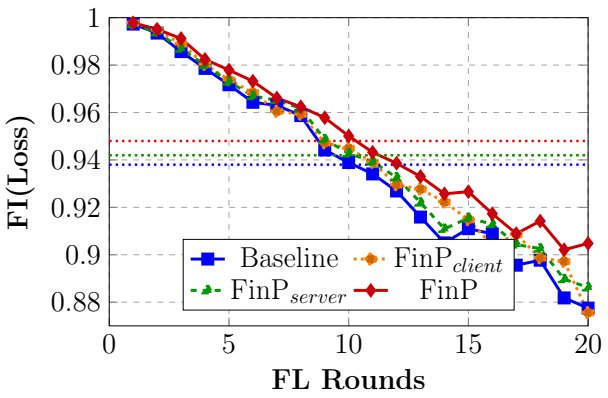


Figure B.5: Disparity of prediction loss FI(Loss) among clients using HAR with ALA.

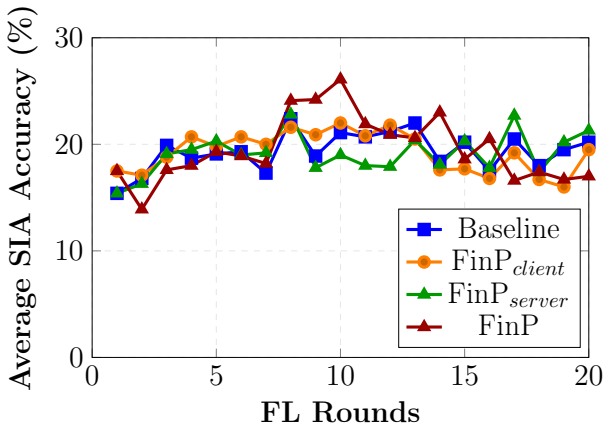


Figure B.6: Average SIA accuracy using HAR dataset with ALA.

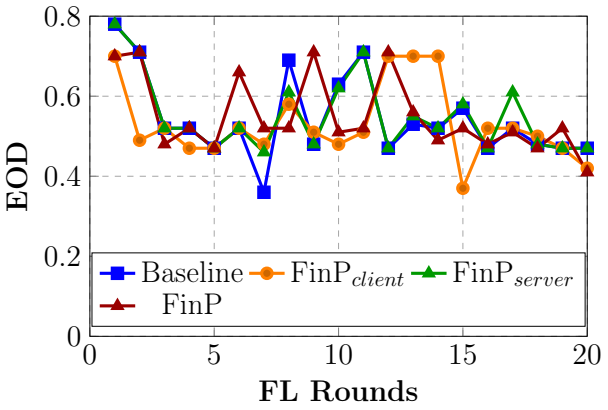
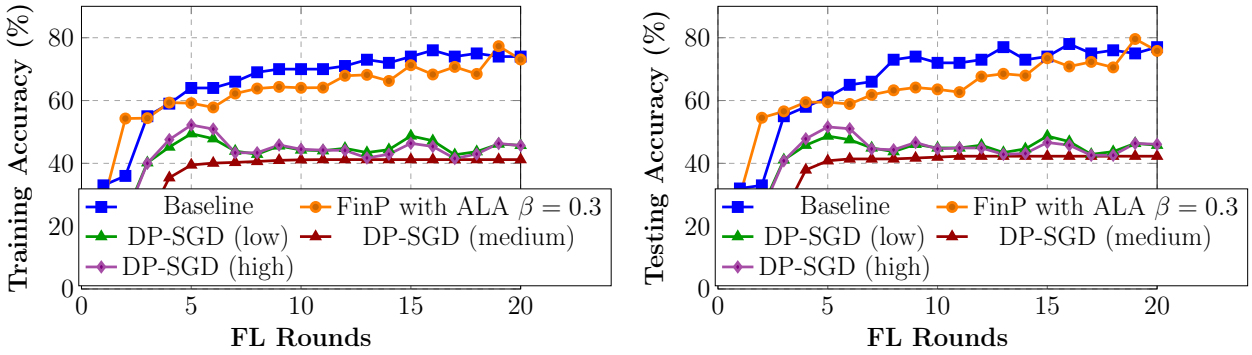


Figure B.7: Equal opportunity difference (EOD) using HAR dataset with ALA.

B.2.3 Differential Privacy (DP) with HAR

Figure B.8 illustrates the training and testing accuracy for HAR when using DP-SGD, while Figures B.9 and B.10 show the FI(SIA) and FI(loss).



(a) Training accuracy.

(b) Testing accuracy.

Figure B.8: Global model classification accuracy using HAR with different levels of noise in DP-SGD.

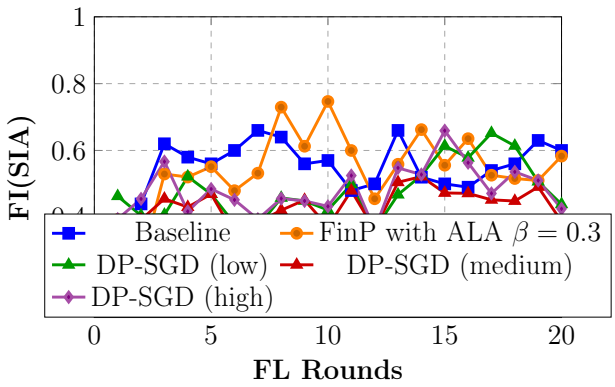


Figure B.9: Disparity of SIA accuracy (FI(SIA)) among clients using HAR dataset with different noise in DP-SGD.

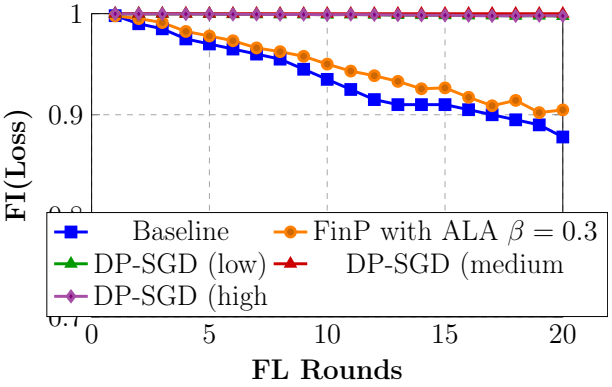


Figure B.10: Disparity of prediction loss FI(Loss) among clients using HAR and different levels of noise in DP-SGD.

B.2.4 Correlation between the top Hessian eigenvalue ($\lambda_{\max}$) and Hessian trace (H_T)

Figure B.13 shows the strong correlation between the top Hessian eigen value ($\lambda_{\max}$) and the Hessian trace (H_T).

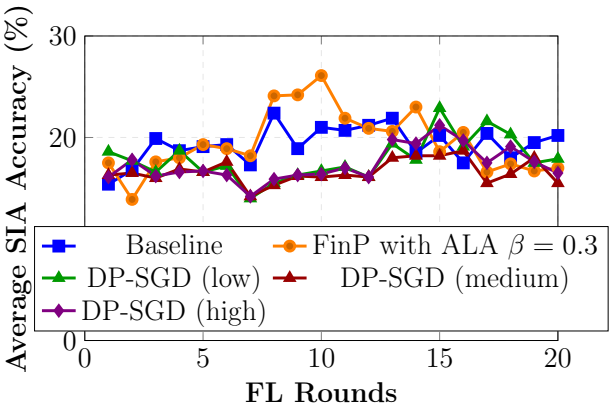


Figure B.11: Average SIA accuracy using HAR dataset with ALA and different noise in DP-SGD.

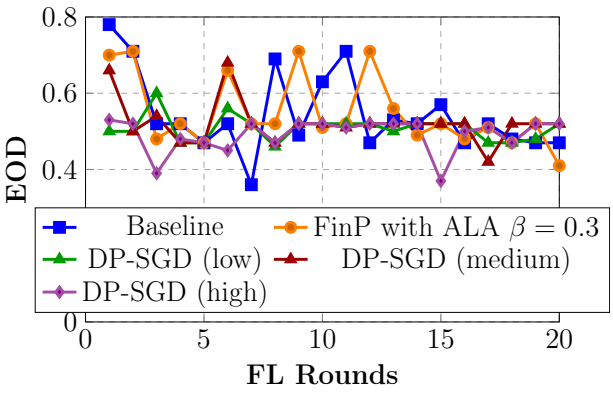


Figure B.12: EOD using HAR dataset with ALA and different noise in DP-SGD.

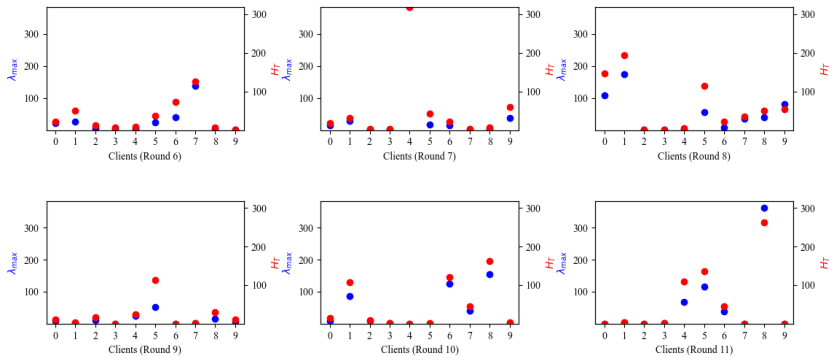


Figure B.13: Strong correlation between the top Hessian eigenvalue ($\lambda_{\max}$, blue dots) and Hessian trace (H_T, red dots) across training rounds in HAR.

Table B.1: Results of CIFAR dataset using ResNet as the local model.

	Accuracy (%)		Privacy Metrics (%))		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA$_{acc}$) ↓	Max (SIA$_{acc}$)↓	CoV(SIA$_{acc}$)/FI(SIA$_{acc}$)	CoV(loss)/FI(Loss)	EOD ↓	Conv. round
Baseline	78.39	76.45	30.86	38.52	**0.33/0.90**	0.67/0.69	0.33	10
FedAlign Mendieta et al. [2022]	70.79	71.87	30.72	38.46	0.34/0.89	0.86/0.58	0.33	14
F*in*P (with PCA)	**80.23**	**78.47**	**10.07**	**10.67**	0.35/0.89	**0.44/0.83**	**0.12**	12

B.3 Experiment Results on CIFAR-10

Complete results for ablation experiments on β, ALA and PCA comparison, and DP-SGD can be found in Table B.2.

Table B.2: Ablation experiments on β. Results of CIFAR dataset using CNN as the local model [Hu et al., 2023].

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics			Efficiency
	Train	Test	Mean (SIA_{acc}) ↓	Max (SIA_{acc})↓	CoV(SIA_{acc})/FI(SIA_{acc})	CoV(loss)/FI(Loss)	EOD ↓	Conv.round
Baseline Hu et al. [2023]	75.62	62.37	40.91	46.70	0.23/0.95	0.46/0.83	0.29	5
FinP								
(β=0.05, PCA)	69.31	59.67	39.51	43.40	0.21/0.96	0.46/0.83	0.27	5
(β=0.1, PCA)	70.45	61.19	39.47	43.70	**0.19/0.96**	0.44/0.84	**0.25**	7
(β=0.3, PCA)	71.17	63.81	**31.85**	**39.90**	0.31/0.90	**0.38/0.87**	0.31	12
(β=0.5, PCA)	10	10	N/A	N/A	N/A	N/A	N/A	N/A
(β=0.05, ALA)	**76.03**	**64.26**	38.62	43.90	0.23/0.95	0.46/0.83	0.29	6
(β=0.1, ALA)	74.64	63.94	37.39	42.50	0.25/0.94	0.44/0.84	0.32	7
(β=0.3, ALA)	71.00	64.09	34.09	41.00	0.34/0.89	0.41/0.86	0.38	11
DP-SGD (ϵ, δ)								
(93, 10^{-5})	43.73	43.59	23.38	24.50	0.52/ 0.79	0.290/0.922	0.37	15
(16, 10^{-5})	42.84	42.62	22.68	24.20	0.45/0.83	0.298/0.918	0.31	15
(5, 10^{-5})	31.52	32.16	21.85	24.60	0.38/0.88	0.271/0.931	0.29	13

B.3.1 Equal Opportunity Difference (EOD) and SIA accuracy

Figures B.14 and B.15 illustrate the average SIA accuracy in CIFAR-10 with CNN across different β, and the EOD results respectively.

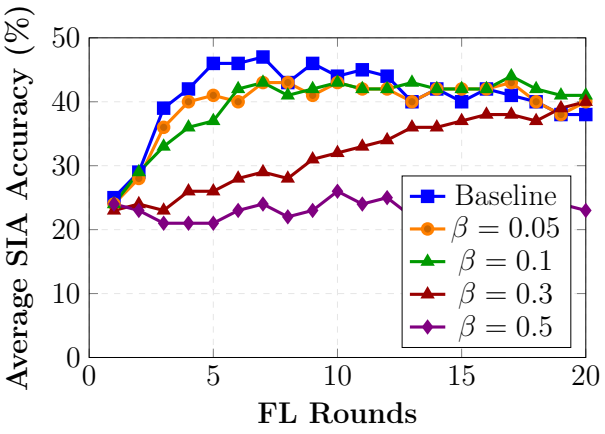


Figure B.14: Average SIA accuracy of Baseline and different β using CIFAR-10 with CNN.

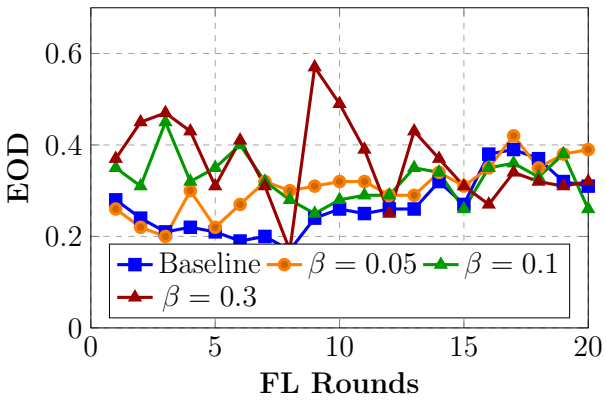
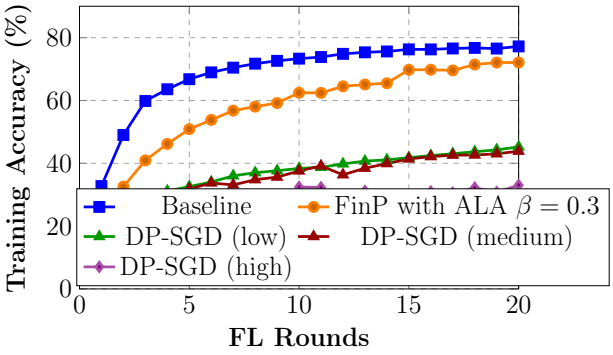


Figure B.15: Equal opportunity difference (EOD) using CIFAR-10 with CNN.

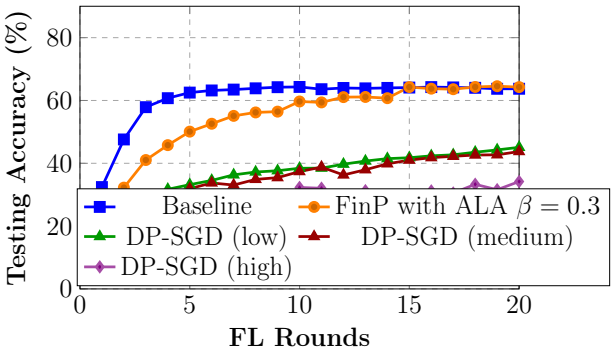
B.3.2 Differential Privacy (DP) with CIFAR-10

Figure B.16 highlights the global model classification accuracy in CIFAR-10 with CNN under different levels of noise in DP-SGD. Figures B.17 and B.18 show the disparity of SIA accuracy (FI(SIA)) and disparity of prediction loss (FI(loss)) among clients using CIFAR-10 with CNN and under different noise in DP-SGD, respectively. Figures B.19 and B.20 illustrate the average SIA accuracy and EOD respectively using ALA in CIFAR-10 with CNN and

different noise in DP-SGD



(a) Training accuracy.



(b) Testing accuracy.

Figure B.16: Global model classification accuracy using CIFAR-10 with CNN and different levels of noise in DP-SGD.

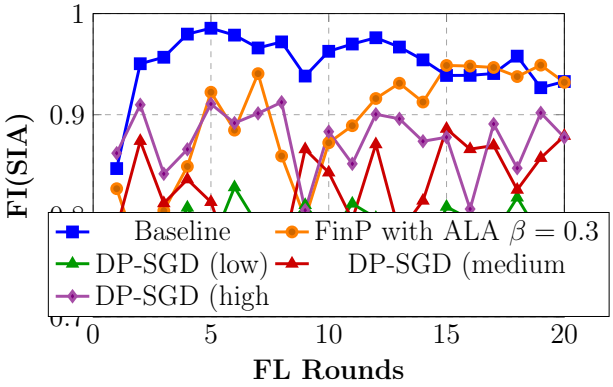


Figure B.17: Disparity of SIA accuracy (FI(SIA)) among clients using CIFAR-10 dataset with CNN and different noise in DP-SGD.

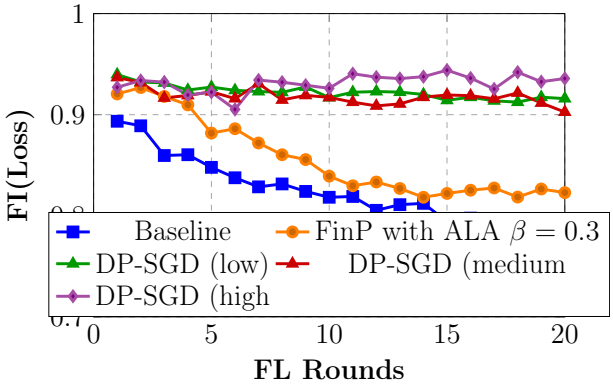


Figure B.18: Disparity of prediction loss FI(Loss) among clients using CIFAR-10 with CNN and different levels of noise in DP-SGD.

B.4 Experiment Results on FEMNIST

B.4.1 Additional Fairness and Disparity Metrics

We utilize fairness notions beyond simple average performance and incorporate metrics that explicitly quantify the distributional disparity among clients. Specifically, we utilize the Mean Absolute Difference (MAD) and Sen's Social Welfare Function.

Mean Absolute Difference (MAD) The Mean Absolute Difference quantifies the average

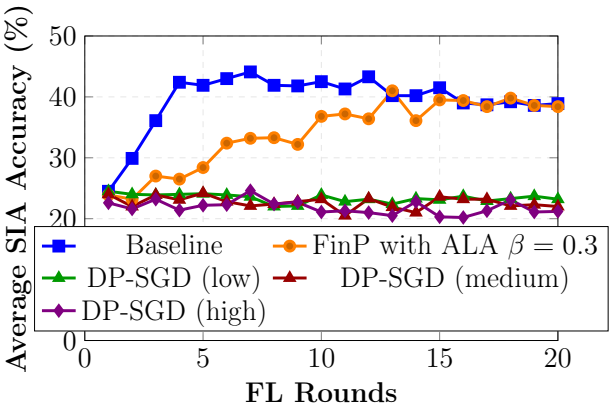


Figure B.19: Average SIA accuracy in CIFAR-10 dataset using ALA with CNN and different noise in DP-SGD.

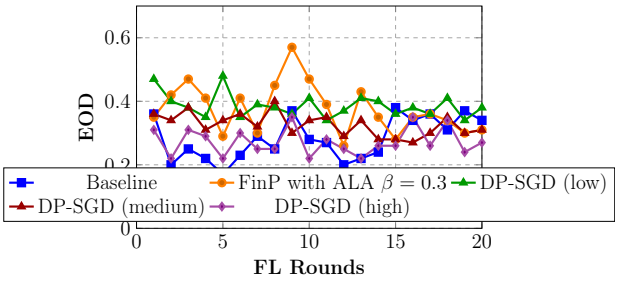


Figure B.20: EOD in CIFAR-10 dataset using ALA with CNN and different noise in DP-SGD.

absolute disparity between all possible pairs of individuals or clients within the system. For a population of size n, where x_i represents the metric evaluated for client i, the MAD is defined as:

$$\text{MAD} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \tag{B.1}$$

Unlike variance, which squares differences and thus heavily weights extreme outliers, MAD provides a linear, highly interpretable measure of overall systemic inequality. Crucially, assuming $x_i \in [0, 1]$, MAD reaches its mathematical upper bound of 0.5 not under a monopoly (where one client holds all utility), but under conditions of *perfect polarization*—where exactly half of the population receives a score of 1 and the other half receives 0. Therefore, MAD is exceptionally well-suited for identifying models that systematically disenfranchise specific clusters or demographics while perfectly serving others and independent of the mean values.

Sen's Social Welfare Function To holistically assess the system, it is necessary to balance overall efficiency with equity. Sen's Social Welfare Function Sen [1997] achieves this by scaling the global average performance (μ) by an inequality penalty. By fully expanding the Gini

coefficient (G) within the standard formulation $W = \mu(1 - G)$, the function is explicitly defined as:

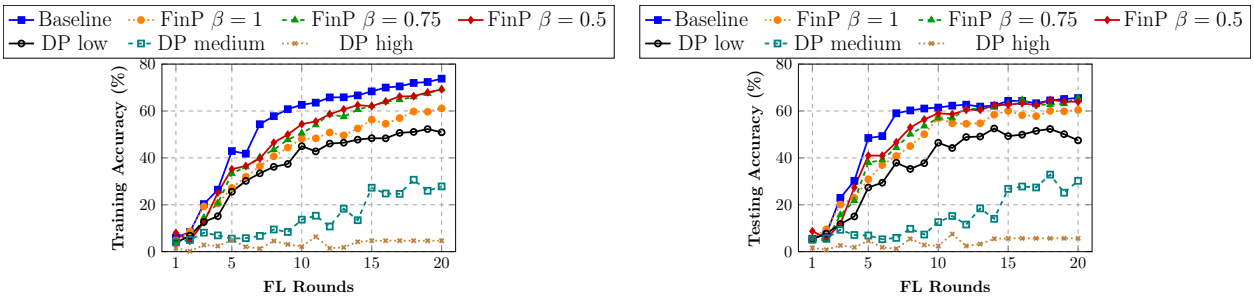
$$W = \mu\,(1 - G) = \mu\left(1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu}\right) \tag{B.2}$$

This formula calculates the egalitarian equivalent performance of the system, assuming the target metric (e.g., utility or predictive accuracy) is one where a higher value is optimal. It bridges the gap between raw average performance and distributional fairness.

- **The Global Mean (μ):** The outer multiplier represents the system's overall average performance. The function inherently rewards pushing this global average as high as possible. *In our context of attack accuracy, where a lower value is optimal, we calculate the μ based on (1-Accuracy$_{attack}$).*

- **The Expanded Disparity Penalty:** The fractional term inside the parentheses is the expanded Gini coefficient. The double summation in the numerator ($\sum\sum|x_i - x_j|$) exhaustively calculates the absolute disparity between every possible pair of clients in the system. The denominator ($2n^2\mu$) normalizes this disparity relative to the total number of pairwise comparisons and the size of the mean itself.

By subtracting this normalized disparity from 1 and multiplying the result by the mean, the function imposes a strict structural penalty on inequality. The system cannot achieve a high Welfare score simply by maximizing performance for a privileged subset of clients (which would inflate μ but also drastically inflate the disparity penalty). Instead, it is mathematically forced to distribute performance as equitably as possible across all clients.



(a) Global model training accuracy.

(b) Global model testing accuracy.

Figure B.21: Global model classification accuracy using FEMNIST dataset. F*in*P (ALA) maintains competitive accuracy with minimal convergence delay.

B.4.2 Hessian Calculation Overhead

To demonstrate the practical viability of deploying FinP in real-world federated environments, we evaluate the computational overhead introduced by the Hessian calculations during our FEMNIST experiments. Because our framework relies primarily on the dominant (top) Hessian eigenvalue instead of full Hessian Matrix, and it does not require exact numerical precision to effectively rank client vulnerability, we can aggressively optimize the solver to significantly reduce the computational burden.

In our experiments, the global neural network architecture has a lightweight memory size of 3.275 MB CNN. To optimize for speed, we configure the eigensolver with highly relaxed convergence constraints: a maximum of 20 iterations for both the eigenvalue solver and the trace estimator (`hessian_eig_max_iter=20`, `hessian_trace_max_iter=20`), alongside a tolerance threshold of 5×10^{-3} (`hessian_tol=5e-3`). Under these lightweight hyperparameters, the dominant Hessian eigenvalue computation requires only 1.82 seconds per client. This minimal overhead confirms that F*in*P can be seamlessly integrated into standard federated training pipelines without imposing a prohibitive computational bottleneck on edge devices.

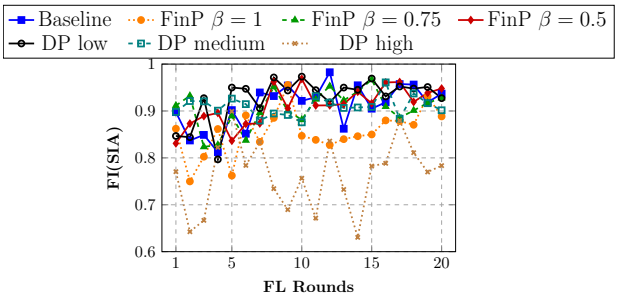


Figure B.22: Disparity of SIA accuracy (FI(SIA)) among clients using FEMNIST dataset with F*in*P with ALA.

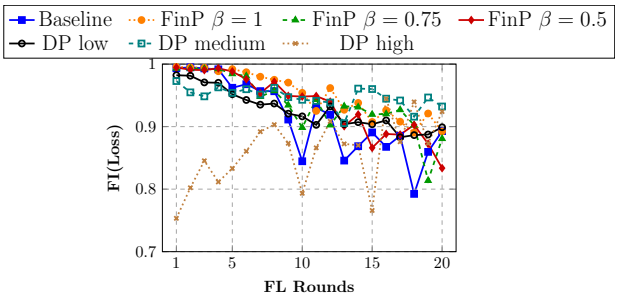


Figure B.23: Disparity of prediction loss FI(loss) among clients using FEMNIST dataset with F*in*P with ALA.

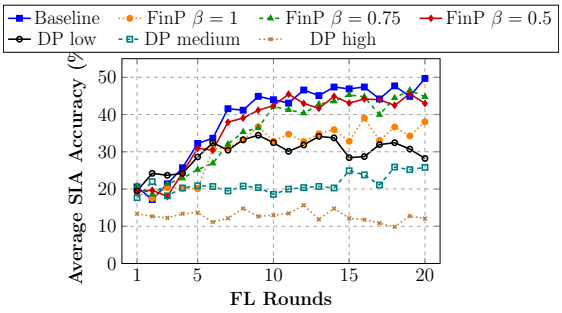


Figure B.24: Average SIA accuracy in FEMNIST with F*in*P with ALA.

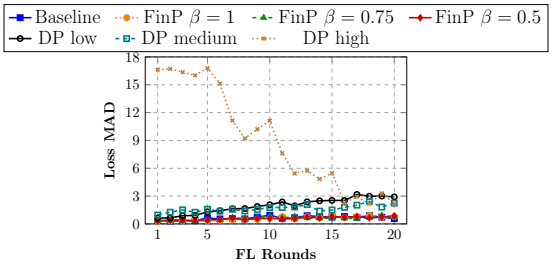


Figure B.25: Loss MAD (mean absolute difference of loss) across clients in FEMNIST with F*in*P with ALA.

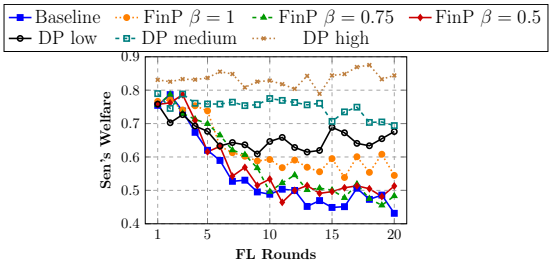


Figure B.26: Sen's welfare across clients in FEMNIST with F*in*P with ALA.

B.5 Detailed Experimental Setup and Results for Per-Round White-Box MIA

To evaluate the vulnerability of individual client updates, we simulate a per-round gradient-matching Membership Inference Attack (MIA). The adversary is defined as the "honest-but-curious" central server, which has access to the global model parameters G_{t-1} and intercepts the raw, unaggregated local weight updates Δw_i from each client i at round t.

B.5.1 Data Partitioning and Threat Model Setup

We partition the overall dataset to ensure strict separation between the adversary's auxiliary knowledge and the final evaluation data. The federated cohort consists of N participating clients.

- **Dummy Client (Client 0):** The adversary injects one sybil client into the federated process. This client participates in aggregation at every round, serving as the "Member" baseline.
- **Victim Clients (Clients 1 to $N-1$):** The genuine participants who is targeted in MIA.
- **Shadow Non-Member Pool ($\mathcal{D}_{shadow}$):** A pool of data (e.g., an isolated writer) held by the adversary, exclusively used to generate "Non-Member" features for classifier training.
- **Evaluation Sets ($\mathcal{D}_{test}$):** For each victim client i, we construct a static evaluation dataset of 100 records: 50 randomly sampled records from client i's local data (Label 1) and 50 records from unseen writer data (Label 0).

B.5.2 Dynamic Feature Generation (Round t)

Because full-participation FL continuously memorizes data over time, static MIA thresholds fail. Instead, the adversary calibrates a dynamically trained Random Forest classifier prior to the aggregation step at each round t. We generate a balanced training set of 60 feature vectors (30 positive, 30 negative) using a bootstrapping approach on the Dummy Client.

For each of the 30 iterations, we generate a paired sample:

1. **Target Sampling:** The adversary samples a member record $x_m \sim$ Client 0 and a non-member record $x_{nm} \sim \mathcal{D}_{shadow}$.

2. **Gradient Extraction:** Using standard Cross-Entropy Loss, the adversary computes the individual gradient footprints g_m and g_{nm} with respect to the current global model G_{t-1}.

3. **Mock Local Update Simulation:** The adversary samples a random 80% subset of Client 0's local dataset, ensuring the inclusion of x_m. A temporary clone of G_{t-1} is trained on this subset using the standard local hyperparameters to yield a mock update Δw_{mock}.

4. **Feature Representation:** To map the high-dimensional parameter space to a robust feature space, the adversary calculates the layer-wise cosine similarity. The positive feature vector is $\cos(g_m, \Delta w_{mock})$, and the negative feature vector is $\cos(g_{nm}, \Delta w_{mock})$.

B.5.3 Classifier Training and Inference

The 60 layer-wise similarity vectors are used to train an Random Forest Classifier. This classifier serves as the decision boundary for round t.

To execute the attack, the adversary intercepts the real uploaded update Δw_i from each victim client $i \in \{1, \ldots, N-1\}$. For each record x_{eval} in the victim's 100-record Evaluation

Table B.3: Results of FEMNIST dataset with MIA attack.

	Accuracy (%)		Privacy Metrics (%)		Fairness Metrics					Efficiency
	Train	Test	Mean (MIA_{acc})↓	Max (MIA_{acc})↓	CoV(MIA_{acc}) /FI(MIA_{acc})	CoV(loss) /FI(Loss)	EOD (MIA)↓	MAD (loss)↓	Welfare (MIA)↑	Conv. round
Baseline Hu et al. [2023]	72.72	65.04	66.63	73.11	0.088/0.99	0.281/0.916	0.40	0.638	0.302	10
F*in*P ($\beta = 1$)	58.65	60.38	**60.53**	**70.56**	0.099/0.990	**0.215/0.949**	0.21	0.606	0.304	12
F*in*P ($\beta = 0.75$)	65.23	62.34	66.57	75.22	**0.089/0.991**	0.229/0.942	**0.18**	**0.574**	0.302	12
F*in*P ($\beta = 0.5$)	**68.03**	**63.76**	86.31	74.78	0.093/0.990	0.257/0.928	0.20	0.608	0.283	14
DP-SGD (ϵ, δ)										
(151, 10^{-5})	49.57	50.10	50.58	58.22	0.038/0.996	0.288/0.920	0.06	2.062	0.483	14
(92, 10^{-5})	27.01	29.01	50.11	52.44	0.038/0.998	0.257/0.938	0.06	2.005	0.488	-
(31, 10^{-5})	2.95	2.50	13.62	16.56	0.041/0.997	0.436/0.838	0.07	10.097	0.493	-

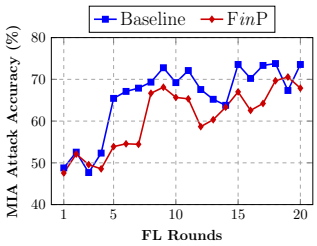


Figure B.27: MIA attack accuracy on FEMNIST with F*in*P with ALA.

Set $\mathcal{D}_{test}$, the adversary computes the gradient g_{eval} on G_{t-1} and extracts the layer-wise cosine similarity against the intercepted Δw_i. The similarity vector is passed to the Random Forest to predict membership. We compute standard classification metrics (Accuracy, Precision, Recall, and ROC-AUC) against the ground-truth labels.

B.5.4 MIA Results

our evaluations demonstrate that the FinP framework inherently provides an effective defense against Membership Inference Attacks (MIAs). Specifically, we observe that MIA accuracy drops by an average of 6.1% under the FinP setup compared to the baseline in Figure B.27 and Table B.3. Although the MIA accuracy disparity across clients are minor (FI = 0.99), F*in*P consistently decreases the disparity of loss (MAD) from 0.638 to 0.574, and decreases the EOD by 0.22 among clients. This empirical reduction highlights that FinP actively mitigates privacy leakage by fostering a more generalized model training process, thereby reducing the localized overfitting and instance-level memorization that gradient-matching MIAs rely upon to succeed.

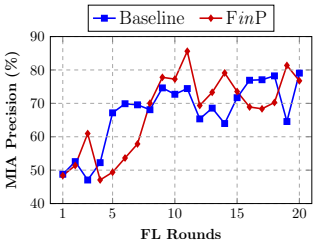


Figure B.28: MIA precision on FEMNIST with F*in*P with ALA.

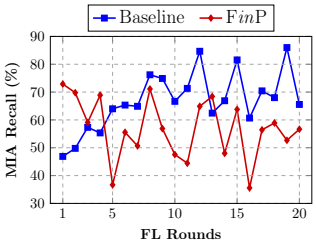


Figure B.29: MIA recall on FEMNIST with F*in*P with ALA.

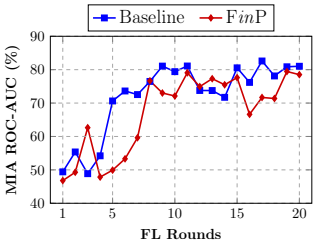


Figure B.30: MIA ROC-AUC on FEMNIST with F*in*P with ALA.

Appendix C

PACIFIC Appendix

C.1 Dataset

C.1.1 Dataset Example

We provide an example from our Psychometrics-based Dataset. Tab. C.1 presents the annotated preference statement, and Tab. C.2 lists the four corresponding answer choices with their labels and OCEAN annotations.

Table C.1: PACIFIC entry example High-E (Preference annotation).

Field	**Trait**	**Value**	**Text**
preference_statement	–	–	I only enjoy live music events where I can be in a general admission pit or a standing-room-only section close to the action.
related_to_personality	–	1	–
score	O	6	–
confidence	O	0.6	–
explanation	O	–	The user's desire to be 'close to the action' in a high-intensity environment like a general admission pit suggests a preference for new and stimulating experiences, which aligns with the 'Adventurousness' facet of high Openness.
confidence_explanation	O	–	The preference for an intense, non-traditional concert experience is a good indicator of adventurousness. However, it's a specific context and may not generalize to other areas of life.
score	C	4	–
confidence	C	0.1	–

Continued on next page

Table C.1: PACIFIC entry example High-E (Preference annotation) (continued).

Field	Trait	Value	Text
explanation	C	–	The statement provides no information about the user's level of organization, discipline, or reliability. The preference for a pit could be seen as either spontaneous (low) or requiring planning (high), making it impossible to assess.
confidence_explanation	C	–	There is no evidence in the provided text to support a score for Conscientiousness in either direction.
score	E	7	–
confidence	E	0.9	–
explanation	E	–	The user's preference for a general admission pit, a crowded, high-energy, and socially dense environment, strongly indicates a high need for stimulation and excitement, which are core facets of Extraversion.
confidence_explanation	E	–	The statement describes a classic preference for an environment sought by individuals high in Extraversion, particularly the 'Excitement-Seeking' and 'Gregariousness' facets. The use of 'only' strengthens this signal.
score	A	4	–
confidence	A	0.1	–
explanation	A	–	The statement describes a preferred environment but does not detail the user's interpersonal style (e.g., cooperative vs. competitive) within that environment. Therefore, agreeableness cannot be assessed.
confidence_explanation	A	–	There is no evidence in the provided text to support a score for Agreeableness in either direction.
score	N	1	–
confidence	N	0.8	–

Continued on next page

Table C.1: PACIFIC entry example High-E (Preference annotation) (continued).

Field	Trait	Value	Text
explanation	N	–	Choosing a general admission pit, an environment with physical discomfort and potential risks (crowds, pushing), over a safer, seated option demonstrates a very low need for comfort and safety, prioritizing the intense experience instead.
confidence_explanation	N	–	The preference is a clear and direct trade-off, sacrificing physical safety and comfort for a more stimulating experience. This provides strong evidence for a low score on this specific scale.

Table C.2: PACIFIC entry example High-E (Choice annotations).

Choice	Field	Trait Score(Confidence)	Text
Conversation query: What's the best way for me to get tickets for the upcoming city music festival?			
Choice 0			
text	–	–	Prioritize getting the 'Front Stage Pit' pass to be right in the middle of the crowd and energy.
related_to_personality	–	1	–
score/conf	O	5 (0.6)	explanation:The user seeks a high-intensity sensory experience by wanting...
score/conf	C	4 (0.2)	explanation: The statement shows a clear goal ('Prioritize getting ...
score/conf	E	7 (1.0)	explanation: The user's explicit desire to be 'right in...
score/conf	A	4 (0.1)	explanation: The choice is focused on a personal sensory goal and does not...

Continued on next page

Table C.2: PACIFIC entry example High-E (Choice annotations) (continued).

Choice	**Field**	**Trait Score(Confidence)**	**Text**
score/conf	N	1 (0.9)	explanation: The choice represents a clear trade-off, sacrificing ...
Choice 1			
text	–	–	Secure a reserved seat in the grandstand for a comfortable and clear, but distant, view.
related_to_personality	–	1	–
score/conf	O	3 (0.6)	explanation: The user chose a practical and conventional option (a reserved seat) ...
score/conf	C	6 (0.8)	explanation: The action to 'secure a reserved seat' implies planning...
score/conf	E	2 (0.8)	explanation: The user opted for a 'distant' view from a grandstand,...
score/conf	A	4 (0.1)	explanation: The choice of seating is a personal preference for comfort and view; ...
score/conf	N	7 (0.9)	explanation: Choosing a 'reserved' and 'comfortable' seat is a risk-averse ...
Choice 2			
text	–	–	Look for tickets in the upper levels, as they are often cheaper and far less crowded.
related_to_personality	–	1	–
score/conf	O	3 (0.6)	explanation: The user prioritizes practical concerns like cost ('cheaper') and...
score/conf	C	6 (0.7)	explanation: The user demonstrates cautiousness and deliberation...
score/conf	E	2 (0.9)	explanation: The user's explicit preference for a 'far less crowded'...

Continued on next page

Table C.2: PACIFIC entry example High-E (Choice annotations) (continued).

Choice	Field	Trait Score(Confidence)	Text
score/conf	A	4 (0.1)	explanation: The choice to seek cheaper, less crowded ...
score/conf	N	6 (0.8)	explanation: The user prioritizes comfort and the avoidance of potential...
Choice 3			
text	–	–	Opt for a VIP package that includes a private, quiet viewing lounge away from the masses.
related_to_personality	–	1	–
score/conf	O	4 (0.3)	explanation: The user is attending a music festival, which can be a novel ...
score/conf	C	6 (0.6)	explanation: Opting for a VIP package implies planning and a desire ...
score/conf	E	2 (0.9)	explanation: The user's explicit desire for a 'private, quiet' space 'away ...
score/conf	A	4 (0.2)	explanation: The statement does not provide clear evidence regarding the user's...
score/conf	N	7 (0.8)	explanation: The choice prioritizes a highly controlled, comfortable, ...

C.1.2 Annotation Criteria

Scoring Criteria for O, C, E, A: To operationalize this, we define the annotation scale based on the *spectrum* of the trait characteristics present in the preference statement:

- *Score 1-3 (Low Trait Expression)*: The preference reflects the characteristics associated with the low end of the trait spectrum (e.g., Routine/Tradition for Openness).
- *Score 4 (Neutral)*: The preference shows no strong directional alignment with the trait.

- *Score 5-7 (High Trait Expression):* The preference reflects the characteristics associated with the high end of the trait spectrum (e.g., Novelty/Complexity for Openness).

Scoring Criteria for N: To operationalize this, we defined the annotation scale for Neuroticism based on the provision of *Safety vs. Efficiency*:

- *Score 1-3 (High Efficiency / High Risk):* The preference prioritizes performance, efficiency, brutal truths, and high stakes. (Target: Low-N Users).
- *Score 4 (Neutral)*: A balance between safety and performance.
- *Score 5-7 (High Safety / High Comfort):* The preference prioritizes guarantees, emotional regulation, and risk avoidance. (Target: High-N Users).

C.1.3 Trait Distribution

The distribution of each trait in PrefEval and PACIFIC(with $\tau = 0.7, 0.8, 0.9$, respectively) is shown in Fig. C.1. PACIFIC provides complete personality-driven preferences, while PrefEval is severely skewed with few preferences in Low-C, High-E, Low-A, and Low-N.

C.1.4 Data construction

To establish a robust benchmark for personality-aware preference alignment, we developed a two-stage pipeline utilizing the Gemini Pro 2.5 API. Our methodology prioritizes high discriminative power by focusing on scenarios where generic model behavior is insufficient.

Personality-Driven Scenario Generation. The first step focused on synthesizing challenging interaction scenarios rooted in distinct personality profiles. For each instance in the dataset, we generated a structured tuple consisting of:

- Personality-Driven Preference: A specific constraint or desire derived from a user persona.
- Preference-Sensitive Query: A user question designed specifically to test the system's adherence to the stated preference.

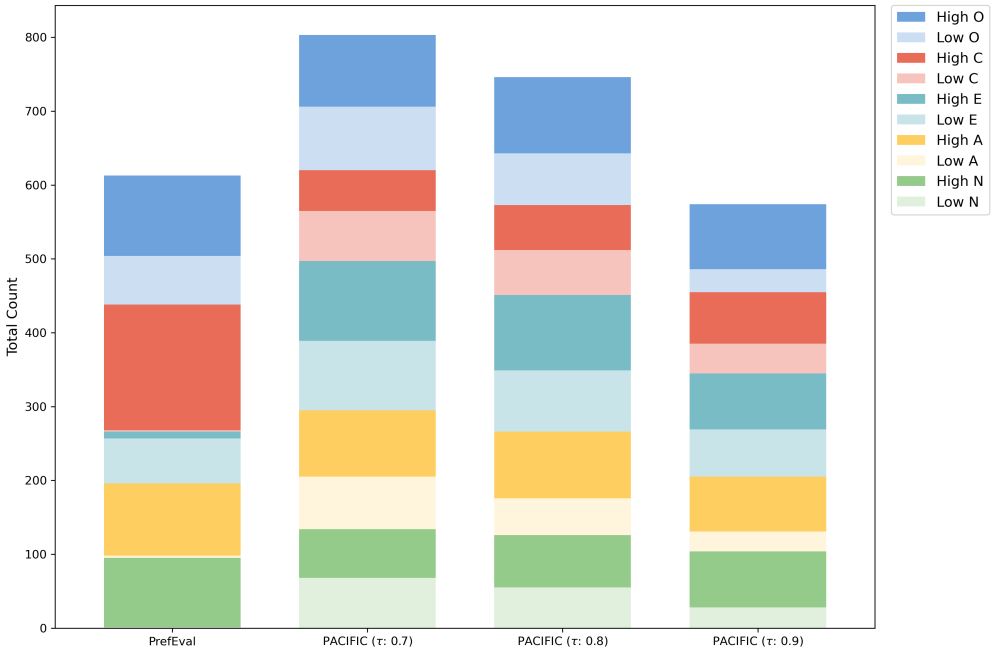


Figure C.1: **Trait distribution overview:** PrefEval exhibits a skewed trait distribution, with limited examples for Low-C, High-E, Low-A, and Low-N. In total, PrefEval has 613 personality-driven preferences out of 1000 preferences with confidence of $\tau = 0.7$. PACIFIC yields 803, 746, 574 personality-driven preferences out of 1200 preferences with confidence of $\tau = 0.7, 0.8, 0.9$ respectively.

- Difficulty Rationale: A concise explanation detailing why this query presents a challenge for standard models.
- Candidate Responses: A set of four response choices, where exactly one option aligns with the user's preference and personality, while the remaining three serve as plausible but misaligned distractors.

A critical constraint in this process was the enforcement of a "High Violation" criterion. We prompted the model to generate queries where a generic or "safe" response would be highly likely to violate the specific user preference. This ensures that the dataset effectively filters for models capable of nuanced personalization rather than generic helpfulness. The prompt is shown in Prompt C.1 and provided in full in this appendix.

We also employ a Confidence Filtering Protocol to derive a high-quality evaluation subset. Let a data sample be defined as S=$\{p_i, q_i, (c_{\text{correct}}, c_{\text{distractor}})\}$, where p is the preference state-

ment, c_{correct} is the ground-truth choice, and $c_{\text{distractor}}$ are the incorrect options. We adopt a confidence threshold $\tau = 0.7$ and apply the following inclusion criteria:

- **Preference Validity:** The preference statement p must reflect the target personality trait (High or Low) with a model confidence score τ.
- **Answer Congruence:** The correct choice c_{correct} must not only be semantically valid but must also explicitly exhibit the same trait polarity as p with confidence τ. This ensures the "correctness" is driven by personality alignment, not just general common sense.
- **Distractor Separation:** To ensure discriminative power, all distractors must fail to meet the target criteria. A distractor is deemed valid only if it either reflects the opposing trait polarity or lacks sufficient confidence (conf$< \tau$) to be confused with a trait-aligned choice.

While the total number of eligible preferences varies with different thresholds, applying a threshold of τ=0.7 yields 803 different preferences and a balanced trait distribution. We set τ=0.7 to filter out uncertain pairs and avoid possible LLM hallucination. Visualizations of trait distributions under other τ values with PrefEval comparison are detailed in Appendix C.1.3.

Dual-Strategy Annotation To guarantee the quality and interpretability of the dataset, we implemented a dual-strategy annotation scheme. This process involved assessing both the latent user preference and the manifest content of each candidate response against the Big Five (OCEAN) personality traits. Prompt can be found in Prompt C.2. The annotation criteria details can be found in Sec.4.2.3.2.

```
You are a helpful assistant.  You are helping a user create scenarios to evaluate if an AI
assistant properly considers the user's stated preferences. You will generate:

Phase 1:  The Scenario
1.  **Personality-driven preference**:...
2.  **Question**:  ...
3.  **Explanation**:  ...

<OCEAN definitions>

Rubric for generating the scenarios:
Please generate such preference question pairs with high Violation probability:        <High
violation definition>
<Hign violation examples>
<Constraints>

Phase 2:  The Options
Given the generated personality-driven preference, question and short explanation from
Phase 1.
Think of exactly 4 possible recommendation options to answer this question in the 'options'
list.
<Dual-Strategy Annotation>

Generate exactly 4 options in the specific order below.
* **Option 1 (First Item in list)**:  The **ADHERING** response.  ...
* **Options 2, 3, 4 (Remaining Items)**:  The **VIOLATING** responses.  ...
```

Prompt C.1: Scenario generation prompt

C.1.5 Experiment result in Prefeval

We report results from the same experiments in the PrefEval Zhao et al. [2025] dataset in Tab. C.3.

C.1.6 Topics Covered in our psychometrics-based dataset

Topics covered in PACIFIC dataset are listed in Table. C.4.

C.1.7 Experimental Results on PACIFIC with Other Models

We evaluate different persona prompting strategies on PACIFIC across multiple models. The results for Llama-3-8B-Instruct, gemini-2.5-pro, and gpt-4o-mini are reported in Tables C.5, C.6, and C.7, respectively.

```
You are a personality assessment expert.  Analyze the following user preference statement
or user behavior and predict their OCEAN (Big Five) personality trait scores.
Input Data

<Preference or Question + 1 choice>
Task:
Assess the user's personality traits using the OCEAN model based on .....
<OCEAN Personality Traits>

<OCEAN definitions>

Assessment Instructions
Step 1:  Relevance Check     Determine if the preference statement is related to personality
traits.  ...
Step 2:  Trait Scoring
<**Trait Score** (1-7 scale integer)>
Step 3:  Assessment Guidelines

<Guidelines>
```

Prompt C.2: personality annotation prompt

Table C.3: Accuracy (%) for different ways of incorporating persona hints into preference prompting from PrefEval Dataset Zhao et al. [2025].

Method	**Variant**	**Overall Acc. (%)**	**Trait-wise Acc. (%)**									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
A. Few-shot	Trait-aligned preferences	82.50	82.50	85.00	–	92.50	75.00	67.50	–	92.50	–	–
B. Few-shot + persona hints	Pref. + pref. trait (GT)	**82.50**	85.00	80.00	–	95.00	67.50	70.00	–	97.50	–	–
	Pref. + pref. & choice traits (GT)	75.83	70.00	80.00	–	95.00	60.00	60.00	–	90.00	–	–
	Traits only (pref. & choices) (GT)	48.33	45.00	65.00	–	45.00	40.00	25.00	–	70.00	–	–
C. Reminder	Instruction-only (no labels)	83.75	85.00	82.50	–	92.50	72.50	70.00	–	97.50	–	–
D. RAG	Pre-trained retriever	74.59	65.00	65.00	–	92.50	65.00	67.50	–	92.50	–	–
	Fine-tuned retriever	80.84	85.00	62.50	–	85.00	87.50	70.00	–	95.00	–	–

Table C.4: Topic taxonomy and example subtopics.

Topic	**Description (examples)**
Home_Cooking	Recipe modifications, kitchen equipment, meal planning, cooking techniques, food storage
Personal_Finance	Bill management, credit score, daily budgeting, banking issues, savings tips
Home_Maintenance	Appliance repairs, plumbing issues, cleaning methods, power problems, HVAC care
Family_Care	Child development, elderly support, family activities, parenting tips, work–life balance
Digital_Services	App troubleshooting, account security, subscription management, device setup, software updates
Weather_Planning	Daily forecasts, event planning, storm preparation, seasonal activities, travel weather
Personal_Documents	ID renewal, document filing, form completion, legal paperwork, record keeping
Shopping_Assistance	Price comparison, product reviews, warranty info, return policies, discount finding
Time_Management	Schedule planning, deadline tracking, calendar organization, task prioritization, routine building
Communication_Help	Email writing, message drafting, call scripts, meeting planning, social media posts
Medical_Care	Symptom checking, medication info, appointment booking, insurance questions, health records
Entertainment_Planning	Event tickets, party planning, holiday activities, group gatherings, weekend ideas
Personal_Growth	Habit formation, goal setting, self-improvement, skill development, career planning
Local_Services	Business hours, service booking, location finding, price inquiries, review checking
Relationship_Advice	Communication tips, conflict resolution, dating guidance, friend issues, family dynamics
Smart_Home	Device control, automation setup, network issues, energy management, security settings
Personal_Safety	Emergency plans, safety precautions, security tips, first aid help, risk assessment
Moving_Relocation	Planning timeline, service booking, address changes, packing tips, set up utilities
Gift_Giving	Gift ideas, occasion planning, budget options, wrapping tips, delivery tracking
Seasonal_Tasks	Holiday planning, weather preparation, wardrobe changes, decoration ideas, activity planning

Table C.5: Accuracy (%) of Llama-3-8B-Instruct under different strategies for incorporating persona hints into preference prompting on PACIFIC.

Method	**Variant**	**Overall Acc. (%)**	**Trait-wise Acc. (%)**									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
Motivation	i) No preferences	33.75	20	50	45	32.5	45	30	35	32.5	17.5	30
	ii) Mixed-trait preferences	29.5	27.5	27.5	17.5	42.5	45	20	30	40	27.5	17.5
	iii) Trait-aligned preferences	**88.75**	85	92.5	92.5	85	87.5	92.5	92.5	95	90	75
	iv) Trait-aligned + 2 noisy preferences	61.75	52.5	35	35	50	60	77.5	87.5	87.5	85	47.5
A. Few-shot	Trait-aligned preferences	**88.75**	85	92.5	92.5	85	87.5	92.5	92.5	95	90	75
B. Few-shot + persona hints	Preference + pref. + trait (GT)	89	97.5	100	95	95	85	95	87.5	95	90	50
	Preference + pref. & choice traits (GT)	32.75	77.5	17.5	72.5	50	5	12.5	22.5	42.5	0	27.5
	Traits only (pref. & choices) (GT)	8.75	2.5	2.5	15	5	0	0	5	10	0	47.5
C. Reminder	Instruction-only (no labels)	74.75	62.5	80	85	82.5	82.5	87.5	70	87.5	62.5	47.5

Table C.6: Accuracy (%) of Gemini-2.5-pro under different strategies for incorporating persona hints into preference prompting on PACIFIC.

Method	**Variant**	**Overall Acc. (%)**	**Trait-wise Acc. (%)**									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
Motivation	i) No preferences	38	20	35	27.5	50	47.5	52.5	32.5	40	22.5	52.5
	ii) Mixed-trait preferences (Total 5 prefs)	61.5	55	72.5	35	67.5	72.5	65	52.5	67.5	62.5	65
	iii) Trait-aligned preferences	99.25	100	100	100	100	100	100	100	97.5	97.5	97.5
	iv) Trait-aligned + 2 noisy preferences	98	97.5	100	100	100	95	100	97.5	100	95	95
	ii) Mixed-trait preferences (Total 25 prefs)	59.5	37.5	70	30	82.5	65	85	35	77.5	62.5	55
A. Few-shot	Trait-aligned preferences	99.25	100	100	100	100	100	100	100	97.5	97.5	97.5
B. Few-shot + persona hints	Preference + pref. + trait (GT)	99.25	100	100	100	100	95	100	100	97.5	100	100
	Preference + pref. & choice traits (GT)	99.5	100	100	100	100	95	100	100	100	100	100
	Traits only (pref. & choices) (GT)	97.75	100	90	100	97.5	92.5	100	100	100	97.5	100
C. Reminder	Instruction-only (no labels)	99.25	100	100	100	100	100	100	100	95	97.5	100

Table C.7: Accuracy (%) of gpt-4o-mini under different strategies for incorporating persona hints into preference prompting on PACIFIC.

Method	**Variant**	**Overall Acc. (%)**	**Trait-wise Acc. (%)**									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
Motivation	i) No preferences	50.5	35	52.5	65	55	65	42.5	75	55	12.5	47.5
	ii) Mixed-trait preferences	63.5	52.5	72.5	55	75	75	55	70	75	52.5	52.5
	iii) Trait-aligned preferences	97.5	97.5	95	100	100	95	100	100	100	95	92.5
	iv) Trait-aligned + 2 noisy preferences	95	100	90	95	95	95	95	100	100	90	90
A. Few-shot	Trait-aligned preferences	97.5	97.5	95	100	100	95	100	100	100	95	92.5
B. Few-shot persona hints	Preference + pref. + trait (GT)	99	100	100	100	100	97.5	97.5	100	97.5	97.5	100
	Preference + pref. & choice traits (GT)	99	100	100	100	100	100	97.5	100	97.5	95	100
	Traits only (pref. & choices) (GT)	97.25	100	95	100	97.5	100	85	100	97.5	97.5	100
C. Reminder	Instruction-only (no labels)	99	100	100	100	100	100	100	100	95	95	100

C.2 Hardware Configuration

All experiments were conducted on a high-performance computing cluster equipped with NVIDIA L40S GPUs. The system runs NVIDIA driver version 580.82.07 with CUDA 13.0 support, and PyTorch was compiled with CUDA 12.1. Each GPU provides 46GB of VRAM (46,068 MiB), enabling efficient processing of large language models. The experiments utilized a single GPU per run during model inference.

C.3 Methods Description

This section summarizes the prompting templates used in our experiments. Across all methods, we present the user's preference context first and place the target question at the end. We also enforce a strict output format: the model must return a single integer in {1,2,3,4}.

C.3.1 Few Shot.

We provide a set of user preference statements as context and ask the model to predict the user's choice for the target question using Prompt C.3.

```
Please analyze the following user preference statements.


<preference>
{
preference_statement_1
preference_statement_2
preference_statement_3
preference_statement_4
preference_statement_5
}
</preference>


Based on these preferences, predict the user's choice for the following question.


<question>
{
conversation_query
Choice 1:  ...
Choice 2:  ...
Choice 3:  ...
Choice 4:  ...
}
</question>


Return a single integer in {1,2,3,4}.
```

Prompt C.3: Few Shot User Preference prompt

```
Please analyze the following user preference statements.

<preference>
{
preference_statement_1
preference_statement_2
preference_statement_3
preference_statement_4
preference_statement_5
noise_preference_statement_1
noise_preference_statement_2
}
</preference>

Based on these preferences, predict the user's choice for the following question.

<question>
{
conversation_query
Choice 1:  ...
Choice 2:  ...
Choice 3:  ...
Choice 4:  ...
}
</question>

Return a single integer in {1,2,3,4}.
```

Prompt C.4: Noise using unrelated preferences

C.3.2 Noise using unrelated preferences.

We provide a set of user preference statements as context and ask the model to predict the user's choice for the target question using Prompt C.4.

C.3.3 Few Shot + Persona Trait Hints

In addition to the preference context, we provide an inferred personality profile (from the preferences) and trait scores for each answer choice using Prompt C.5. The model is instructed to select the choice that best matches both the user's preferences and the provided profile.

C.3.4 RAG

Table C.8: Accuracy (%) for pretrained retriever and fine-tuned retriever

Method	Variant	Overall Acc. (%)	Trait-wise Acc. (%)									
			$\mathbf{O}^H$	$\mathbf{C}^H$	$\mathbf{E}^H$	$\mathbf{A}^H$	$\mathbf{N}^H$	$\mathbf{O}^L$	$\mathbf{C}^L$	$\mathbf{E}^L$	$\mathbf{A}^L$	$\mathbf{N}^L$
RAG	pretrained retriever	30.25	27.50	22.50	10.00	22.50	45.00	32.50	35.00	42.50	37.50	27.50
	fine-tuned retriever	43.00	32.50	30.00	32.50	37.50	52.50	40.00	55.00	62.50	52.50	35.00

```
You are a decision engine that selects the best-matching choice using Big Five (OCEAN)
personality traits.

OCEAN definitions:
{
'O': 'Openness to Experience:  creativity, curiosity, willingness to try new ideas',
...
'N': 'Neuroticism:  emotional stability (low) vs. anxiety and reactivity (high)'
}

<preference>
{
preference_statement_1
...
preference_statement_5
}
</preference>

User personality profile (inferred from the preferences above):
<user_profile>
{
trait:  O, level:  High
}
</user_profile>

<question>
{
conversation_query
Choice 1:  ...
...
Choice 4:  ...
}
</question>

Predicted trait scores for each choice (1-7 per trait):
<choices>
{
Choice 1:  {O: , C: , E: , A: , N: }
...
Choice 4:  {O: , C: , E: , A: , N: }
}
</choices>

Task:  Choose the single best option (1-4) that best matches the user profile.
Output (STRICT): Return ONLY one integer in {1,2,3,4}.
```

Prompt C.5: Few Shot + Persona Trait Hints Prompt

Preferences + Trait Labels of Preferences and Choices. We further expose the model's intermediate signals by providing (i) an OCEAN-based user profile inferred from the preference statements and (ii) predicted OCEAN trait scores for each candidate answer choice shown in Prompt C.6. The model is then instructed to select the option whose trait direction and scores best align with the inferred user profile.

```
You are a decision engine that selects the best-matching choice using Big Five (OCEAN)
personality traits.

OCEAN definitions:
{
'O': 'Openness:  creativity, curiosity, willingness to try new ideas',
...
}

<preferences>
{
preference_statements
}
</preferences>

User personality profile (inferred from the preferences above):
<user_profile>
{
trait_direction:  {O: High, C: Low, E: High, A: High, N: Low}
}
</user_profile>

<question>
{
conversation_query
Choices:  ...
}
</question>

Predicted trait scores for each choice (1-7 per trait; include direction when available):
<choices>
{
option_1:  {O-high:  7, A-high:  6, N-low:  2}
...
option_4:  {O-low:  2, E-high:  6, N-high:  7}
}
</choices>

Task:  Choose the single best option (1-4) that most closely matches the user profile.
Comparison rules (apply in order):
- Prefer options with the same trait directions (high/low) as the user profile.
- If multiple options match, choose the one with the most matching traits and closest
scores.
- If none match clearly, choose the closest overall match given the preferences and
choices.

Output (STRICT): Return ONLY one integer in {1,2,3,4}.  No extra text.
```

Prompt C.6: Preferences + Trait Labels of Preferences and Choices Prompt.

```
You are a decision engine that selects the best-matching choice using Big Five (OCEAN)
personality traits.

OCEAN definitions:
{
'O': 'Openness:  creativity, curiosity, willingness to try new ideas',
...
'N': 'Neuroticism:  emotional stability (low) vs. anxiety/reactivity (high)'
}

User personality profile:
<user_profile>
{
trait_direction:  {O: High, C: Low, E: High, A: High, N: Low}
}
</user_profile>

You will be given 4 options.  Each option includes predicted OCEAN trait scores (1-7).
Select the single best option (1-4) whose trait profile most closely matches the user
profile.

<choices>
{
option_1:  {O-high:  7, A-high:  6, N-low:  2}
...
option_4:  {O-low:  2, E-high:  6, N-high:  7}
}
</choices>

Comparison rules (apply in order):
- Prefer options with the same trait directions (high/low) as the user profile.
- If multiple options match, choose the one with the most matching traits and closest
scores.
- If none match clearly, choose the closest overall match.

Output (STRICT): Return ONLY one integer in {1,2,3,4}.  No extra text.
```

Prompt C.7: Preferences' and Choices' Trait Labels Only Prompt.

Preferences' and Choices' Trait Labels Only. To isolate the effect of trait signals, we remove the raw preference statements and provide only (i) an inferred OCEAN user profile and (ii) predicted OCEAN trait scores for each answer option using Prompt C.7. The model is instructed to select the option whose trait directions (high/low) and scores best match the user profile.

```
Please analyze the following user preference statements.

<preference>
{
preference_statement_1
...
preference_statement_5
}
</preference>

Based on these preferences, predict the user's choice for the following question.

Reminder:  Consider the user's personality implied by their preference statements and use
it to guide your prediction.
<question>
{
conversation_query
Choice 1:  ...
...
Choice 4:  ...
}
</question>

Return a single integer in {1,2,3,4}.
```

Prompt C.8: Prompt with personality reminder

C.3.5 Reminder

We use the same prompt as in the few-shot setting, but add a short reminder encouraging the model to consider personality signals implied by the preferences.

```
Reminder:  Consider the user's personality implied by their preference statements and use
it to guide your prediction.
```

Hence, the full prompt is shown in Prompt C.8:

C.4 RAG Fine-tuning

For fine-tuning, we follow the `biencoder_local` configuration with the following hyperparameters: batch size = 1, dev batch size = 16, Adam $\epsilon = 10^{-8}$, Adam betas $(0.9, 0.999)$, weight decay = 0.0, and learning rate = 5×10^{-7}. We fine-tune for 28 epochs until convergence, which takes approximately one hour using the same hardware configuration in Appendix C.2.

C.5 Persona Analysis

We evaluate **32 persona profiles** derived from the Big Five (OCEAN) traits. Since each trait can take one of two levels (*high* or *low*), the full set contains $2^5 = 32$ distinct personas. Table. C.9 lists all personas and their corresponding trait configurations. Using these personas, we compare performance across profiles by measuring **accuracy on trait-aligned preference prediction** for each persona. The results are summarized below in Figure C.2.

Table C.9: The 32 OCEAN persona profiles (H=High, L=Low). Columns index personas (0–31); rows denote traits.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
O	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
C	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L
E	H	H	H	H	L	L	L	L	H	H	H	H	L	L	L	L	H	H	H	H	L	L	L	L	H	H	H	H	L	L	L	L
A	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L
N	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L

C.6 Cross-Validation Results

To further assess the stability of our benchmark, we conduct a 3-fold cross-validation study on our dataset. As shown in Tab. C.10, the mean performance remains consistent across folds and aligns with the main results reported in the manuscript. The standard deviation across folds is small for all settings, remaining below 2%, which confirms the stability of the

benchmark.

Preference context	Fold 0	Fold 1	Fold 2	Mean	Std
i) Zero-shot	21.77%	25.75%	22.73%	23.41%	1.69%
ii) Mixed Traits	28.41%	24.25%	25.00%	25.89%	1.81%
iii) Aligned Traits	61.99%	66.42%	64.02%	64.14%	1.81%
iv) Contaminated Aligned	56.46%	59.70%	59.85%	58.67%	1.56%

Table C.10: 3-fold cross-validation results on our dataset. The low standard deviation across folds indicates that the benchmark results are stable.

C.7 Memory in Agentic AI for user modeling

Prevalent memory architectures, often reduced to RAG-based retrieval, treat user modeling as a naive data storage task. However, this approach could collapse under longitudinal interaction. For scenarios spanning one year of user-LLM interaction, attempting to manage raw history is computationally infeasible and creates reasoning difficulties and noise, making it impossible to effectively retrieve or reason over user preferences amidst the complexity. We argue that effective agentic memory requires shifting from lossless storage to semantic abstraction. It is promising to utilize personality traits as a compression layer; much like a human companion who predicts needs based on "knowing" a person rather than recalling every specific event, personality serves as a dense, high-signal proxy that drives user behavior, preference, and decision-making, allowing agents to accurately predict intent without the overhead of exhaustive retrieval.

To test this, we scaled the number of mixed-trait preferences provided in the LLM's context from 5 to 25. Yet, our experiments revealed that expanding the preference context actually degraded LLM accuracy from 61.75% to 59.5% (Appendix C.1.7, Tab. C.6). This demonstrates that simply feeding an LLM more raw interaction history does not equate to better user modeling; rather, the model becomes overwhelmed by competing signals and noise.

C.8 Limitation & Future work

Our study has several limitations that open avenues for future work.

First, we simplify the personality spectrum into a binary High/Low classification. In reality, Big-Five traits are continuous, and individuals sharing the same High or Low direction can still differ in nuanced ways that a binary label cannot capture. In future work, we plan to adopt a finer-grained representation, using the full 7-point Likert scores directly to better reflect the continuous variation in human personality.

Second, our evaluation focuses on multiple-choice question answering. We plan to extend PACIFIC to open-ended generative tasks and compare performance against the QA baseline, assessing whether trait-aligned personalization transfers to free-form generation.

Third, we assume that personality traits are independent of cultural and linguistic context. In practice, the way traits manifest in behavior and preferences may vary across societies, cultures, and languages. We leave a systematic investigation of these cross-cultural and cross-lingual effects to future work.

Impact Statement

PACIFIC studies how personality trait information (based on the Big Five/OCEAN framework) can improve an LLM's ability to follow user preferences in multiple-choice question answering, and proposes retrieval-based methods for selecting trait-aligned preference statements when annotations are unavailable. The intended positive impact is to advance research on controllable and user-aligned generation, which may help build assistants that are more consistent with a user's stated goals and reduce frustrating or irrelevant responses.

We view PACIFIC primarily as a measurement and method-development contribution, and we encourage responsible use. We do not foresee specific ethical concerns or negative soci-

Figure C.2: Accuracy (%) by persona group (32 total). Each panel corresponds to one group (Idx 0–31) and shows the overall accuracy plus per-trait accuracies for the Big Five (O, C, E, A, N). Bar colors encode traits (Overall in magenta with white hatch; O/C/E/A/N in distinct colors), and titles indicate each group's high/low trait configuration.

etal consequences arising from PACIFIC beyond those commonly associated with improving the capability and controllability of machine learning models. To support transparency and reproducibility, we plan to release our dataset and code, and we follow standard research practices and a code of conduct for responsible use. Overall, we expect PACIFIC to contribute to more personalized and helpful LLM systems in the future, improving user experience and better assisting people in everyday tasks.

Appendix D

Preference Aggregation Appendix

D.1 Experiment Configurations and Hyperparameters

Our experimental setup begins with supervised fine-tuning (SFT) as outlined in Table D.1. We use the Gemma-2-2b-it model as our base, employing LoRA adaptation with rank 16 for efficient parameter updates. The SFT training utilizes a cosine learning rate scheduler with warmup and is conducted for a single epoch to establish our baseline model.

As summarized in Table D.2, both the policy and value models in PPO are initialized from the SFT model. During training, we employ two distinct prompt formats for evaluation: a preference probability prediction task requiring models to assign probability scores to all options, and a preference ranking task requiring complete ordinal ranking from most to least preferred (see Figures 5.7 and 5.8). Our implementation builds upon the Hugging Face TRL library von Werra et al. [2022]. All experiments were conducted on 3 nodes, each equipped with A100 GPUs, Intel(R) Xeon(R) Gold 6326 CPUs @ 2.90GHz, and 256GB RAM.

[1]Rewards are whitened over each rollout before PPO updates.

Table D.1: SFT configuration and hyperparameters.

Hyperparameter	**Value**
Model	
Base model	`google/gemma-2-2b-it`
Precision	BF16
Data / Task	
Train/valid split	80/20%
Max sequence length	500 (include prompt and response)
LoRA Adapter	
Rank (r)	16
Alpha	32
Dropout	0.05
Optimization	
Batch size (per device)	16
Gradient accumulation steps	4
Learning rate	5×10^{-5}
Scheduler	cosine
Warmup steps	150
Weight decay	0.01
Training	
Epochs	1

Table D.2: PPO configuration and hyperparameters (policy and value models initialized from the SFT model).

Hyperparameter	**Value**
General	
Policy model	Gemma 2 SFT model
Value model	Gemma 2 SFT model
Model / Quantization	
Quantization	4-bit (`nf4`, double-quant = True)
Compute dtype	BF16
Attention implementation	eager
LoRA (PEFT)	
Rank (r)	32
Alpha	32
Dropout	0.05
Optimization	
Per-device train batch size	4
Gradient accumulation steps	24
Learning rate	1×10^{-5}
Optimizer	AdamW
Weight decay	0.0
Scheduler	linear
PPO Trainer	
PPO epochs	2
Mini-batches	8
Per-device eval batch size	32
Response length	42
Temperature	0.6
KL coefficient	0.05
Clip range	0.2
Clip range (value)	0.2
Value loss coefficient (v_f)	0.2
Discount factor (γ)	1.0
GAE lambda (λ)	0.95
Reward whitening	Per rollout (before PPO update)[1]

	Group-level fairness							Q/A-level fairness						
Model	Binary	Borda	KendallTau	KL	Cosine	Wass.	JS	Binary	Borda	KendallTau	KL	Cosine	Wass.	JS
Base model	0.999398	0.999869	0.997414	0.999080	0.999983	0.999997	0.999990	0.936377	0.892326	0.891544	0.941767	0.996680	0.999219	0.998310
SFT model	0.999634	0.999932	0.997102	0.998037	0.999966	1.000000	0.999983	0.941832	0.889914	0.893297	0.936904	0.996414	0.999193	0.998214
RL model	0.999016	0.999823	0.998064	0.992068	0.999974	0.999992	0.999928	0.918107	0.896839	0.900015	0.908564	0.999312	0.999420	0.998625

Table D.3: Fairness scores across reward functions (7 metrics each) JS denotes JensenShannonReward.

Model	Binary	Borda	KendallTau	KL	Cosine	Wass.	JS
Base model	0.237087	0.408545	0.205106	-0.561043	0.739418	0.777263	0.686737
SFT model	0.243285	0.396608	0.197353	-0.503460	0.761378	0.784242	0.703906
RL model	0.296488	0.477538	0.334573	-0.196769	0.902256	0.868361	0.820557

Table D.4: Alignment scores across reward functions. JS denotes JensenShannonReward.