UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Reliable Foundations for Interactive Machine Learning

Permalink

https://escholarship.org/uc/item/7mg6k75r

ISBN

9798270242190

Author

Joren, Hailey

Publication Date

2025-12-20

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Reliable Foundations for Interactive Machine Learning

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Hailey Lynn Joren

Committee in charge:

Professor Berk Ustun, Chair
Professor Taylor Berg-Kirkpatrick
Professor Sanjoy Dasgupta
Professor Julian McAuley

2025

Copyright

Hailey Lynn Joren, 2025

All rights reserved.

The Dissertation of Hailey Lynn Joren is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

TABLE OF CONTENTS

Dissertation Approval Page iii

Table of Contents iv

List of Figures vi

List of Tables vii

Acknowledgements viii

Vita ix

Abstract of the Dissertation x

Chapter 1 Participatory Personalization in Classification 1
1.1 Introduction 1
1.2 Participatory Systems 5
1.3 Learning Participatory Systems 9
1.4 Experiments 13
1.4.1 Setup 14
1.4.2 Discussion 15
1.5 Concluding Remarks 19

Chapter 2 Classification with Conceptual Safeguards 21
2.1 Introduction 21
2.2 Framework 25
2.3 Methodology 29
2.3.1 Uncertainty Propagation 29
2.3.2 Confirmation 31
2.4 Experiments 32
2.4.1 Setup 33
2.4.2 Results 34
2.5 Concluding Remarks 37

Chapter 3 Sufficient Context: A New Lens on Retrieval Augmented Generation Systems 39
3.1 Introduction 39
3.2 Related Work 41
3.3 Sufficient Context 43
3.3.1 Definition of Sufficient Context 44
3.3.2 Sufficient Context AutoRater 45
3.4 A New Lens on RAG Performance 47
3.4.1 Do Benchmark Datasets Have High Sufficient Context? 48
3.4.2 Initial Findings Based on Sufficient Context 49

3.4.3 Qualitatively analyzing responses with insufficient context 51
3.5 Techniques to Reduce Hallucinations with RAG 53
3.5.1 Selective RAG Using Sufficient Context Signal 53
3.5.2 Fine-tuning 55
3.6 Conclusion 56

Appendix A Participatory Personalization 58
A.1 Supporting Material for Section 1.3 58
A.1.1 Enumeration Routine for Algorithm 1 58
A.1.2 Assignment Routine for Algorithm 1 59
A.1.3 Pruning Routine for Algorithm 1 60
A.1.4 Greedy Induction of Sequential Reporting Interface 61
A.2 Description of Datasets used in Section 1.4 – Experiments 62
A.3 Results for Different Model Classes and Prediction Tasks 65
A.3.1 Logistic Regression for Ranking (AUC) 65
A.3.2 Random Forests for Decision-Making (Error) 67
A.3.3 Random Forests for Ranking (AUC) 68
A.4 Supporting Material for Performance Profiles 68

Appendix B Conceptual Safeguards 70
B.1 Omitted Proofs 70
B.2 Supporting Material for Experiments 71
B.2.1 Datasets 71
B.2.2 Additional Experimental Results 73
B.2.3 Multiclass Classification Tasks 75

Appendix C Sufficient Context 76
C.1 Supporting Experiment Information 76
C.1.1 Models 76
C.1.2 Fine-tuning Settings 77
C.1.3 Datasets 77
C.2 Additional Results 79
C.2.1 Fine-tuning Full Results 79
C.2.2 Performance Breakdown by Sufficient Context 80
C.2.3 Comparison of QA Evaluation Metrics 84
C.3 Prompts 86
C.3.1 Sufficient Context Autorater Prompt 86
C.3.2 FLAMe Prompt 87
C.3.3 LLMEval Prompt 88
C.3.4 Dataset Question Answer Prompts 89
C.4 Fine-Tuning and RAG Prompts for Mistral 91

Bibliography 93

LIST OF FIGURES

Figure 1.1. Classification task where participation improves accuracy and minimizes data use. 2

Figure 1.2. Participatory systems for a personalized classification task with group attributes 7

Figure 1.3. Performance profile of participatory systems for the `saps` dataset for each intersectional group in the `saps` dataset 10

Figure 1.4. Participatory systems for the `saps` dataset. 19

Figure 2.1. Conceptual safeguard to detect melanoma from an image of a skin lesion. We consider a model that estimates the probabilities of *m* concepts: `Dotted`, `Pigmented` ... `IrregularVasc` 24

Figure 2.2. We evaluate the performance of classification models that can abstain using an accuracy-coverage curve. 28

Figure 3.1. New insights into RAG systems by looking at whether instances have sufficient context. 42

Figure 3.2. We compare the % of instances that our autorater labels as sufficient across datasets, either with the first 10k, 6k, or 2k tokens of the provided sources. 48

Figure 3.3. Model Performance on Datasets Stratified by Sufficient Context. 50

Figure 3.4. Selective Generation: Coverage vs. Selective Accuracy. 56

Figure C.1. Correct, hallucination, and abstention fractions across models for dataset FreshQA, stratified by sufficient context. 82

Figure C.2. Correct, hallucination, and abstention fractions across models for dataset HotpotQA, stratified by sufficient context. 83

LIST OF TABLES

Table 1.1. Metrics used to evaluate performance, data use, and consent of personalized models and systems. 15

Table 1.2. Participatory systems and personalized models for all datasets. 16

Table 2.1. Accuracy coverage curves for all methods on all datasets. 35

Table 2.2. Coverage at specific thresholds τ for a confirmation budget of 20%. 36

Table 3.1. Sufficient Context AutoRater. 46

Table 3.2. Qualitative Analysis of Correct Answer & Insufficient Context. 52

Table A.1. Datasets used to fit clinical prediction models in Section 1.4 62

Table A.2. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc** 66

Table A.3. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test error** 67

Table A.4. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc** 68

Table B.1. Coverage across abstention thresholds τ with confirmation budget 10% . . . 73

Table B.2. Coverage across abstention thresholds τ with confirmation budget 50% . . . 74

Table B.3. Coverage vs. accuracy for all methods on multiclass classification tasks. . . 75

Table C.1. Fine-tuned (FT) Mistral 3 7B Instruct 80

Table C.2. Performance Analysis of RAG Systems Using Human-Annotated Sufficient Context Labels. 81

Table C.3. Comparison of evaluation metrics across models and datasets. 84

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Berk Ustun for his guidance as the chair of my committee and my advisor. I am also grateful to my talented research collaborators, as well as my spouse, Taylor Joren, whose love and support made this possible.

Chapter 1, in full, is a reprint of the material as it appears in Neural Information Processing Systems 36 (NeurIPS 2023). Hailey Joren; Chirag Nagpal; Katherine A. Heller; and Berk Ustun. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in The Twelfth International Conference on Learning Representations (ICLR 2024). Hailey Joren; Charles Marx and Berk Ustun. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in The Thirteenth International Conference on Learning Representations (ICLR 2025) Hailey Joren; Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. The dissertation author was the primary investigator and author of this paper.

VITA

2019 Bachelor of Arts in Computer Science, Cum Laude in Field, Harvard College

2025 Doctor of Philosophy in Computer Science, University of California San Diego

ABSTRACT OF THE DISSERTATION

Reliable Foundations for Interactive Machine Learning

by

Hailey Lynn Joren

Doctor of Philosophy in Computer Science

University of California San Diego, 2025

Professor Berk Ustun, Chair

Modern machine learning systems often struggle to balance performance with individual agency, safety, and transparency, particularly in high-stakes applications. To address these challenges, this dissertation presents frameworks for mitigating hallucinations in large language models, formalizing abstention mechanisms in classification, and operationalizing user agency in personalized prediction tasks. Each chapter will detail its theoretical foundations, practical algorithms, and empirical benefits. Together, these contributions outline a principled path toward interactive, self-aware prediction systems that are both safer and more user centric.

Chapter 1

Participatory Personalization in Classification

1.1 Introduction

Machine learning models routinely assign predictions to *people* – be it to screen a patient for a mental illness Kessler et al. [2005], their risk of mortality in an ICU Pollard et al. [2018], or their likelihood of responding to treatment Abajian et al. [2018]. Many models in such applications are designed to target heterogeneous subpopulations using features that explicitly encode personal information. Typically, models are *personalized* with categorical attributes that define groups [i.e., "categorization" as per Fan and Poole, 2006]. In medicine, for example, clinical prediction models use *group attributes* that are *protected* (e.g., `sex` in the ASCVD Score for cardiovascular disease), *sensitive* (e.g., `HIV_status` in the VA COVID-19 Mortality Score), *self-reported* (e.g., `alcohol_use` in the HAS-BLED Score for Major Bleeding Risk), or *costly* to acquire (e.g., `leukocytosis` in the Alvarado Appendicitis Score).

Individuals expect the right to opt out of providing personal data and the ability to understand how it will be used [see, e.g., personal data guidelines in GDPR, OECD privacy guidelines European Parliament and of the Council, 2016, OECD, 2013]. In many contexts, personalized models do not provide such functionality: individuals cannot opt out of reporting data used to personalize their predictions, nor tell if it would improve their predictions. At the same time, practitioners assume that data available for training will be available at inference

Group	Data		Personalized		Generic		Traditional Personalization groups receive predictions from h			Minimal Participatory System groups opt into predictions from h or h_0		
							Model	Data Use	Gain	Model	Data Use	Gain
g	n_g^+	n_g^-	h	$R_g(h)$	h_0	$R_g(h_0)$		r	$\Delta R_g(h,h_0)$		r	$\Delta R_g(h,h_0)$
female, old	0	24	+	24	−	0	h	female, old	−24	h_0	∅	0
female, young	25	0	+	0	−	25	h	female, young	25	h	female, young	25
male, old	25	0	+	0	−	25	h	male, old	25	h	male, old	25
male, young	0	27	−	0	−	0	h	male, young	0	h_0	∅	0
Total	50	51		24		50			26			50

Figure 1.1. Classification task where participation improves accuracy and minimizes data use. We consider a dataset that has no features, two group attributes $\mathscr{G} = \texttt{sex} \times \texttt{age}$, $n^- = 51$ negative examples and $n^+ = 50$ positive examples. Here, the best personalized linear model $h : \mathscr{X} \times \mathscr{G} \to \mathscr{Y}$ with a one-hot encoding of $\mathscr{G}$ makes 24 mistakes, and the best generic model $h_0 : \mathscr{X} \times \mathscr{Y}$ makes 50 mistakes as it predicts the majority class $(-)$. Under traditional personalization, individuals report group membership to receive personalized predictions from h. As shown, personalization benefits the population as a whole by reducing overall error from 50 to 24 ($\Delta R_g(h,h_0) = 26$). However, personalization has a detrimental effect on $[\texttt{female},\texttt{old}]$, who receive less accurate predictions from the personalized model ($\Delta R_g(h,h_0) = -24$), and no effect on $[\texttt{male},\texttt{young}]$ who receive the same predictions from the personalized and generic models ($\Delta R_g(h,h_0) = 0$). In a minimal participatory system, individuals *opt in* to personalization, choosing to receive predictions from h or h_0. Here, individuals in groups $[\texttt{female},\texttt{old}]$ and $[\texttt{male},\texttt{young}]$ opt out of personalization, leading to an overall error of 0 ($\Delta R_g(h,h_0) = 50$) and a reduction in unnecessary data collection ($\varnothing$).

time. In practice, this assumption has led to a proliferation of models that use information that individuals may be unwilling or unable to report at prediction time [see e.g., the Denver HIV Risk Score Haukoos et al., 2012, which asks patients to report `age`, `gender`, `race`, and `sexual_practices`]. In tasks where individuals self-report, they may not voluntarily report information that could improve their predictions or may report incorrect information.

The broader need to facilitate and inform consent in personalized prediction tasks stems from the fact that personalization may not improve performance for each group that reports personal data [Suriyakumar et al., 2023]. In practice, a personalized model can perform *worse* or the same as a *generic model* fit without personal information for a group with specific characteristics. Such models violate the implicit promise of personalization as individuals report personal information without receiving a tailored performance gain in return. These instances of "worsenalization" are prevalent, hard to detect, and hard to resolve [see Suriyakumar et al., 2023, Paes et al., 2022]. However, they would be resolved if individuals could opt out of personalization

and understand its expected gains (see Fig. 1.1).

This work introduces a family of classification models that operationalize informed consent called *participatory systems*. Participatory systems *facilitate consent* by allowing individuals to report personal information at prediction time. Moreover, they *inform consent* by showing how reporting personal information will change their predictions. Models that facilitate consent operate as markets in which individuals trade personal information for performance gains. This work seeks to develop systems that perform as well as possible both when individuals opt-in – to incentivize voluntary reporting – and when they opt out – to safeguard against abstention. Our main contributions include:

1. We present a variety of participatory systems that provide opportunities for individuals to make informed decisions about data provision. Each system ensures that individuals who opt into personalization will receive the most accurate possible predictions.
2. We develop a model-agnostic algorithm to learn participatory systems. Our approach can produce a variety of systems that promote participation and minimize data use in deployment.
3. We conduct a comprehensive study of participatory systems in real-world clinical prediction tasks. The results show how our approach can facilitate and inform consent in a way that improves performance and minimizes data use.
4. We provide a Python library to build and evaluate participatory systems.

Related Work

Participatory systems support modern principles of responsible data use articulated in OECD privacy guidelines [OECD, 2013], the GDPR [European Parliament and of the Council, 2016], and the California Consumer Privacy Act [Bukaty, 2019]. These include: *informed consent*, i.e., that data should be collected with the data subject's consent; and *collection limitation*, i.e., that data collected should be restricted to only what is necessary. These principles stem from extensive work on the right to data privacy [Kaminski, 2019]. They are motivated, in part, by research showing that individuals care deeply about their ability to control personal

data [Bansal et al., 2010, Anderson and Agarwal, 2011, Auxier et al., 2019] but differ considerably in their desire or capacity to share it [see e.g. Awad and Krishnan, 2006, Ortlieb and Garner, 2016, Campbell and Kracaw, 1980, Chemmanur, 1993, Auer et al., 2008, Lundberg et al., 2019, Arellano et al., 2018]. Our proposed systems let decision subjects report personal data in exchange for performance, which is aligned with principles articulated in recent work on data privacy Biega et al. [2020], Shanmugam et al. [2022] and related to work in designing incentive-compatible prediction functions Eliaz and Spiegler [2022].

We consider models that are personalized with categorical attributes that encode personal characteristics [i.e., "categorization" rather than "individualization" as per Fan and Poole, 2006]. Modern techniques for learning with categorical attributes [see e.g., Agresti, 2018, Steyerberg et al., 2019] use them to improve performance at a population level – e.g., by accounting for higher-order interaction effects [Bien et al., 2013, Lim and Hastie, 2015, Vaughan et al., 2020] or recursive partitioning [Elmachtoub et al., 2018, Biggs et al., 2020, Bertsimas et al., 2019, Bertsimas and Kallus, 2020]. Our methods can be used to achieve these goals in tasks where models use features that are optional or costly to acquire [see e.g., Atan et al., 2016, Auer et al., 2008, Yu et al., 2009, Tran and Fioretto, 2023].

Our work is related to algorithmic fairness in that we seek to improve model performance at a group level. Recent work shows that personalization with group attributes does not uniformly improve performance and can reduce accuracy at a group level [see Suriyakumar et al., 2023, Paes et al., 2022, Ustun et al., 2019]. Our systems can safeguard against such instances of "worsenalization" by informing users of the gains in reporting and allowing them to opt out of reporting. This line of broad work complements research on preference-based fairness [Zafar et al., 2017, Ustun et al., 2019, Kim et al., 2019, Viviano and Bradic, 2020, Do et al., 2021], on ensuring fairness across complex group structures [Kearns et al., 2018, Hébert-Johnson et al., 2018, Globus-Harris et al., 2022], and promoting privacy across subpopulations [Tran et al., 2021, Biega et al., 2020].

1.2 Participatory Systems

We consider a classification task where we personalize a model with categorical attributes. We start with a dataset $\{(\boldsymbol{x}_i, y_i, \boldsymbol{g}_i)\}_{i=1}^n$ where each example consists of a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, a label $y_i \in \mathscr{Y}$, and a vector of m categorical attributes $\boldsymbol{g}_i = [g_{i,1}, \ldots, g_{i,m}] \in \mathscr{G}_1 \times \ldots \times \mathscr{G}_m = \mathscr{G}$. We refer to $\mathscr{G}$ as *group attributes*, and to $\boldsymbol{g}_i$ as the *group membership* of person i. We use $n_{\boldsymbol{g}} := |\{i \,|\boldsymbol{g}_i = \boldsymbol{g}\}|$ denote the size of group $\boldsymbol{g}$, and use $|\mathscr{G}_k|$ to denote the number of categories for group attribute k.

We use the dataset to train a personalized model $h : \mathscr{X} \times \mathscr{G} \to \mathscr{Y}$ via empirical risk minimization with a loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$. Given a model h, we denote its empirical risk and true risk as $\hat{R}(h)$ and $R(h)$, respectively, and evaluate model performance at the group level. We denote the empirical risk and true risk of a model h on group $\boldsymbol{g} \in \mathscr{G}$ as

$$R_{\boldsymbol{g}}(h(\cdot, \boldsymbol{g})) := \mathbb{E}()[\ell(h(\cdot, \boldsymbol{g}), y)], \qquad \hat{R}_{\boldsymbol{g}}(h(\cdot, \boldsymbol{g})) := \frac{1}{n_{\boldsymbol{g}}} \sum_{i:\boldsymbol{g}_i = \boldsymbol{g}} \ell(h(\cdot, \boldsymbol{g}), y_i).$$

We consider tasks where every individual prefers more accurate predictions.

Assumption 1. Given models h and h', individuals in group $\boldsymbol{g}$ prefer h to h' when $R_{\boldsymbol{g}}(h) < R_{\boldsymbol{g}}(h')$.

Assumption 1 holds in settings where every individual prefers more accurate predictions – e.g., clinical prediction tasks such as screening or diagnosing illnesses Ustun et al. [2017], Struck et al. [2017]. It does not hold in applications where some individuals prefer predictions that may be inaccurate – e.g., such as predicting the risk of organ failure for a transplant [see e.g., Paulus and Kent, 2020, for other "polar" clinical applications].

Operationalizing Consent

We consider models where individuals consent to personalization by deciding whether or not to report their group attributes at prediction time. We let $\varnothing$ denote an attribute that was not reported, and let $\boldsymbol{r}_i = [r_{i,1}, \ldots, r_{i,k}] \in \mathscr{R} \subseteq \mathscr{G} \times \boldsymbol{\varnothing}$. For example, a person with $\boldsymbol{g}_i =$

$[\texttt{female}, \texttt{HIV} = +]$ would report $\boldsymbol{r}_i = [\texttt{female}, \varnothing]$ if they only disclose `sex`, and would report $\boldsymbol{r}_i = \boldsymbol{\varnothing} := [\varnothing, \varnothing]$ if they opt out of reporting entirely.

We associate each model with a set of *reporting options* $\mathscr{R}$. A traditional model, which requires each person to report group attributes, has $\mathscr{R} = \mathscr{G}$. A model where each person could report any subset of group attributes has $\mathscr{R} = \mathscr{G} \times \boldsymbol{\varnothing}$. We represent individual decisions to opt into personalization at prediction time through a *reporting interface* defined below.

Definition 1. Given a personalized classification task with group attributes $\mathscr{G}$, a *reporting interface* is a tree T whose nodes represent attributes reported at prediction time. The tree is rooted at $\text{root}(T) = [\varnothing, \ldots, \varnothing]$ and branches as a person reports personal attributes. Given a node $\boldsymbol{r}$, we denote its parent as $\text{pa}(\boldsymbol{r})$. Each parent-child pair represents a *reporting decision*, and the height of the tree represents the maximum number of reporting decisions.

Definition 2. Given a personalized classification task with group attributes $\mathscr{G}$, a *participatory system* with reporting interface T is a prediction model $f_T : \mathscr{X} \times \mathscr{R} \to \mathscr{Y}$ that obeys the following properties:

(P1) *Baseline Performance*: Opting out of personalization entirely guarantees the expected performance from a *generic model* trained without group attributes $h_0 \in \arg\min_{h \in \mathscr{H}} R(h)$.

$$R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{\varnothing})) = R_{\boldsymbol{r}}(h_0) \text{ for all reporting groups } \boldsymbol{r} \in \mathscr{R}.$$

(P2) *Incentive Compatibility*: Opting into personalization improves expected performance

$$R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r})) < R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r}'))$$
$$\text{for all nested reporting groups } \boldsymbol{r}, \boldsymbol{r}' \in \mathscr{G} \times \boldsymbol{\varnothing}$$
$$\text{such that } \boldsymbol{r}' = \text{pa}(\boldsymbol{r}).$$

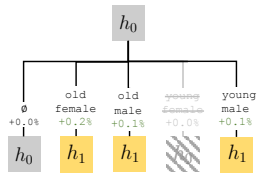
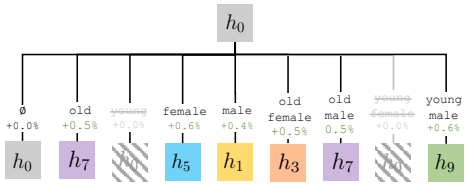
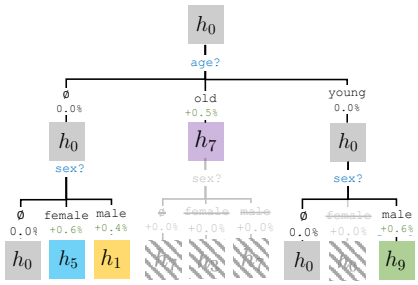
Interface	Example	Description
Minimal		*Minimal systems* let individuals opt out of personalization from an existing personalized model h. Individuals who opt out receive predictions from a generic model h_0 trained without group attributes.
Flat		*Flat systems* support *partial personalization* by allowing each person to report any subset of group attributes. Thus, a person with $\boldsymbol{g}_i = [\texttt{old}, \texttt{female}]$ can report $\boldsymbol{r}_i = [\texttt{old}, \varnothing]$. These systems can improve performance by using a distinct model to assign personalized predictions to each reporting group.
Sequential		*Sequential systems* support partial personalization while allowing individuals to report one attribute at a time. This interface is well-suited to inform consent, as users can make reporting decisions by comparing two models simultaneously. They are also well-suited for personalization tasks with information that must be acquired at prediction time (e.g., the outcome of a test result).

Figure 1.2. Participatory systems for a personalized classification task with group attributes $\texttt{sex} \times \texttt{age} = [\texttt{male}, \texttt{female}] \times [\texttt{old}, \texttt{young}]$. Each system allows a person to opt out of personalization by reporting $\varnothing$ and informs their choice by showing the expected gains of personalization (e.g., $\texttt{+0.2\%}$ gain in accuracy). Systems minimize data use by removing reporting options that do not improve accuracy (see grey-striped boxes). Here, $[\texttt{young}, \texttt{female}]$ is pruned in all systems as it leads to a gain $\leq 0.0\%$.

Here, the *Baseline Performance* property ensures that individuals who choose not to share personal information receive the performance of a generic model – i.e., the most accurate model that could be trained without this information. This property also ensures individuals retain the ability to opt out of personalization – i.e., $\boldsymbol{\varnothing} \in \mathscr{R}$. The *Incentive Compatibility* property ensures that personalization will improve expected performance – i.e., when individuals report personal data, the system can effectively leverage that data to deliver more accurate predictions in expectation. Together, these properties lead to data minimization, as systems that obey these properties will not request data from a reporting group when it will not lead to an improvement in expected performance.

On Data Minimization via Imputation

An alternative approach to allow individuals to opt out of reporting personal information at prediction time is to impute their group membership. Imputation allows individuals to opt out of personalization but does not guarantee the accuracy of their predictions. As a result, individuals who opt out of personalization by reporting $\boldsymbol{r} = \varnothing$ may receive a less accurate prediction than they would receive from a generic model. In the best-case scenario where we could perfectly impute group membership, a group might be assigned better predictions from a generic model (see Fig. 1.1). In the worst case, imputation may be incorrect, leading to even more inaccurate predictions than those of the generic or personalized model. We highlight these effects on real-world datasets in our experiments in Section 1.4.

Characterizing System Performance

One of the key differences between traditional models and participatory systems is that their performance depends on individual reporting decisions. In what follows, we characterize the performance under a general model of individual disclosure. Given a participatory system f_T, we assume that each individual reports personal information to maximize an individual utility function of the form:

$$u_i(\boldsymbol{r}; f_T) = b_i(\boldsymbol{r}; f_T) - \gamma_i \boldsymbol{r} \tag{1.1}$$

Here, $\gamma_i \cdot$ and $b_i(\cdot)$ denote the cost and benefit that individual i receives from reporting $\boldsymbol{r}$ to f_T respectively. We assume that individuals incur no cost when they do not report any attributes such that $\gamma_i \varnothing = 0$, and incur costs that increase monotonically with information disclosed such that $\gamma_i \boldsymbol{r} \leq \gamma_i \boldsymbol{r}'$ for $\boldsymbol{r} \subseteq \boldsymbol{r}'$. We assume that benefits increase monotonically with expected gains in true risk so that $R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r})) < R_{\boldsymbol{r}}(f_T(\cdot, \boldsymbol{r}')) \implies b_i(\boldsymbol{r}, f_T) > b_i(\boldsymbol{r}', f_T)$.

In Fig. 1.3, we show how the system performance for each reporting group can change with respect to participation when we simulate individual disclosure decisions from a model that satisfies the assumptions listed above. When a personalized model h requires individuals to

report information that reduces performance as in Fig. 1.1, individuals incur a cost of disclosure without receiving a benefit in return. In such cases, individuals who interact with a minimal system would opt out of worsenalization and receive more accurate predictions from a generic model, thereby improving the overall performance of the system.

We observe that the maximum utility that each individual can receive from a participatory system can only increase as we add more reporting options. Thus, flat and sequential systems should exhibit better performance than a minimal system.

Given a participatory system f_T with reporting options $\mathscr{R}$, a participatory system $f_{T'}$ with more reporting options $\mathscr{R}' \supseteq \mathscr{R}$ can only improve performance, – i.e., $R(f_{T'}) \leq R(f_T)$. Similarly, the system with more reporting options can only improve utility, – i.e., $u_i(\boldsymbol{r}; f_{T'}) \geq u_i(\boldsymbol{r}; f_T)$ for all individuals i.

1.3 Learning Participatory Systems

This section describes a model-agnostic algorithm to learn participatory systems that ensures incentive compatibility and baseline performance in deployment. We outline our procedure in Algorithm 1 to learn the three kinds of participatory systems in Fig. 1.2. The procedure takes as input a pool of candidate models $\mathscr{M}$, a dataset for model assignment $\mathscr{D}^{\text{assign}}$, and a dataset for pruning $\mathscr{D}^{\text{prune}}$. It outputs a collection of participatory systems that obey the properties described in Definition 2 on test data. The procedure combines three routines to (1) generate viable reporting interfaces (Line 1); (2) assign models over the interface (Line 3); (3) prune the system to limit unnecessary data collection (Line 4). We present complete procedures for each routine in Appendix A.1 and discuss them below.

Model Pool

Our procedure takes as input a *pool of candidate models* $\mathscr{M}$ to assign over a reporting interface. At a minimum, every pool should contain two models: a personalized model h for individuals who opt into personalization, and a generic model h_0 for individuals who opt out of

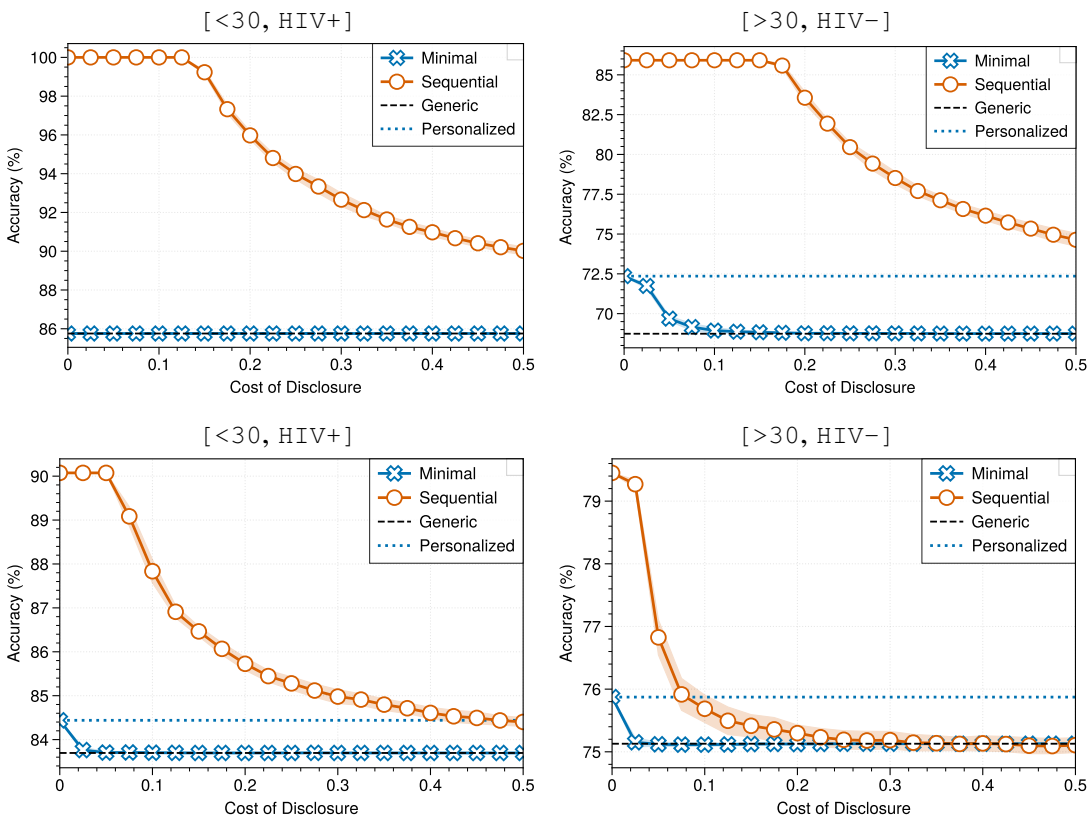


Figure 1.3. Performance profile of participatory systems for the `saps` dataset for each intersectional group in the `saps` dataset. We plot out-of-sample performance for different levels of participation in the target population. We control participation by varying the reporting cost in a simulated model of individual disclosure. As shown, minimal and sequential systems outperform a generic model at a group level regardless of participation. In regimes where the cost of disclosure is low, participation is high. Consequently, a minimal system will achieve the same performance as a personalized model, and a sequential system will achieve the performance of the component model for each subgroup. We provide details and results in Appendix A.4.

Algorithm 1. Learning Participatory Systems

Input: $\mathscr{M} : \{h : \mathscr{X} \times \mathscr{G} \to \mathscr{Y}\}$ *pool of candidate models*

Input: $\mathscr{D}^{\text{assign}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{assign}}}$ *assignment dataset*

Input: $\mathscr{D}^{\text{prune}} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, y_i)\}_{i=1}^{n^{\text{prune}}}$ *pruning dataset*

1: $\mathbb{T} \leftarrow \mathsf{ViableTrees}(\mathscr{G}, \mathscr{D}^{\text{assign}})$ $|\mathbb{T}| = 1$ *for minimal & flat systems*

2: **for** $T \in \mathbb{T}$ **do**

3: $\quad T \leftarrow \mathsf{AssignModels}(T, \mathscr{M}, \mathscr{D}^{\text{assign}})$ *assign models*

4: $\quad T \leftarrow \mathsf{PruneLeaves}(T, \mathscr{D}^{\text{prune}})$ *prune models*

5: **end for**

Output $\mathbb{T}$, collection of participatory systems

personalization. A single personalized model can perform unreliably across reporting groups due to differences in the data distribution or trade-offs between groups. Using a pool of models safeguards against these effects by drawing on models from different model classes that have been personalized using different techniques for each reporting group. By default, we include models trained specifically on the data for each reporting group, as such models can perform well on heterogeneous subgroups Ustun et al. [2019], Suriyakumar et al. [2023].

Enumerating Interfaces

We call the ViableTrees routine in Line 1 to enumerate *viable* reporting interfaces. We only call this routine for sequential systems since minimal and flat systems use a single reporting interface that is known a priori. ViableTrees takes as input a group attributes $\mathscr{G}$ and a dataset $\mathscr{D}^{\text{assign}}$. It returns all m-ary trees that obey constraints on sample size and reporting (e.g., users who report `male` should report `age` before `HIV`). By default, we only generate trees so that we have sufficient data to estimate gains at each node of the reporting interface[1]. In general, ViableTrees scales to tasks with ≤ 8 group attributes. Beyond this limit, one can reduce the enumeration size by specifying ordering constraints or a threshold number of trees to enumerate before stopping. For a task with three binary group attributes, $\mathbb{T}$ contains 24 3-ary trees of depth 3. Given a complete ordering of all 3 group attributes, however, $\mathbb{T}$ would have 1 tree. We can also consider a greedy algorithm (see Appendix A.1.4), which may be practical for large-scale problems.

Model Assignment

We assign each reporting group a model using the AssignModels routine in Line 3. Given a reporting group $\boldsymbol{r}$, we consider all models that could use any subset of group attributes in $\boldsymbol{r}$. Thus, a group that reports `age` and `sex` could be assigned predictions from a model that requires `age`, `sex`, both, or neither. This implies that we can always assign the generic model to any

[1]For example, trees whose leaves contain at least one positive sample, one negative sample, and $n_{\boldsymbol{r}} \geq d+1$ samples to avoid overfitting

reporting group, ensuring that the model at each node performs as well as the generic model on out-of-sample data (i.e., *baseline performance* in Definition 2).

Pruning Reporting Options

AssignModels may output trees that violate incentive compatibility by requesting personal information that fails to improve performance. This can happen when the routine assigns a model that performs equally well to nested reporting groups – see, e.g., Fig. 1.2 where the Flat system assigns h_0 to $[\texttt{female}, \varnothing]$ and $[\texttt{female}, \texttt{young}]$.

We can avoid requesting data from reporting groups in such cases by calling the Prune routine in Line 4. This routine takes as input a participatory system f_T and a pruning dataset $\mathscr{D}^{\text{prune}}$ and outputs a system $f_{T'}$ with a pruned interface $T' \subseteq T$. The routine uses a bottom-up pruning procedure that calls a one-sided hypothesis test at each node:

$$H_0 : \Delta_{\boldsymbol{r}}(\boldsymbol{r}, \text{pa}(\boldsymbol{r})) \leq 0 \qquad H_A : \Delta_{\boldsymbol{r}}(\boldsymbol{r}, \text{pa}(\boldsymbol{r})) > 0$$

The test checks if each reporting group $\boldsymbol{r}$ receives more accurate predictions from the personalized model assigned to its current node or $\boldsymbol{r}$ its parent $\text{pa}(\boldsymbol{r})$. Here, H_0 assumes a reporting group prefers the parent model. Thus, we reject H_0 when we can reliably tell that $f_T(\cdot, \boldsymbol{r})$ performs better for $\boldsymbol{r}$ on the pruning dataset. The exact test should be chosen based on the performance metric for the underlying prediction task. In general, we can use a bootstrap hypothesis test [DiCiccio and Efron, 1996] and draw on more powerful tests for salient performance metrics [e.g., Dietterich, 1998, DeLong et al., 1988, Sun and Xu, 2014, for accuracy and AUC].

On Computation

Our approach provides several options to moderate the computation cost of training a pool of models. For example, we can train only two models and build a minimal system. Alternatively, we can also build a flat or sequential system using a limited number of models in the pool. In practice, the primary bottleneck when building participatory systems is *data*

rather than *compute*. Given a finite sample dataset, we are limited in the number of categorical attributes used for personalization. This is because we require a minimum number of samples for each intersectional group to train a personalized model and evaluate its performance. Given that the number of intersectional groups increases exponentially with each attribute, we quickly enter a regime where we cannot reliably evaluate model performance for assignment and pruning [see Paes et al., 2022].

On Customization

Our procedure allows practitioners to learn systems for prediction tasks by specifying the performance metric used in assignment and pruning. A suitable performance metric should represent the gains we would show users (e.g., error for a diagnosis, AUC for triage, ECE for risk assessment). Using a pool of models allows practitioners to optimize performance across groups, which translates to gains at the population level. For sequential systems, the procedure outputs all configurations, allowing practitioners to choose between systems based on criteria not known at training time. For example, one can swap the trees to use a system that always requests HIV status last. By default, we select the configuration that minimizes data collection across groups, such that the ordering of attributes results leads to the most significant number of data requests pruned.

1.4 Experiments

We benchmark participatory systems on real-world clinical prediction tasks. Our goal is to evaluate these approaches in terms of performance, data usage, and consent in applications where individuals have a low reporting cost. We include code to reproduce these results in an Python library.

1.4.1 Setup

We consider six classification tasks for clinical decision support where we personalize a model with group attributes that are protected or sensitive (see Table 1.2 and Appendix A.2). Each task pertains to an application where we expect individuals to have a low cost of reporting and to report personal information when there is any expected gain. This is because the information used for personalization is readily available, relevant to the prediction task, and likely to be disclosed given legal protections related to the confidentiality of health data [U.S. Congress, 1996, Bansal et al., 2010, Anderson and Agarwal, 2011]. One exception is `cardio_eicu` and `cardio_mimic`, which are personalized based on race and ethnicity. [2] We split each dataset into a test sample (20% for evaluating out-of-sample performance) and a training sample (80% for training, pruning, assignment, and estimating gains to show users). We train three kinds of personalized models for each dataset:

- *Static*: These models are personalized using a one-hot encoding of group attributes (1Hot), and a one-hot encoding of intersectional groups (mHot)
- *Imputed*: These are variants of static models where we impute the group membership for each person (KNN-1Hot, KNN-mHot). In practice, personalized systems with imputation will range between the performance for these systems and the performance of 1Hot and mHot.
- *Participatory*: These are participatory systems built using our approach. These include Minimal, a minimal system built from 1Hot and its generic counterpart; and Flat and Seq, flat and sequential systems built from 1Hot, mHot and their generic counterparts.

We train all models – personalized models and the components of participatory systems – from the same model class and evaluate them using the metrics in Table 1.1. We repeat the experiments four times, varying the model class (logistic regression, random forests) and the target performance metric (error rate for decision-making tasks, AUC for ranking tasks) to

[2] The use of race in clinical risk scores should be approached with caution Vyas et al. [2020]; participatory systems offer one way to safeguard against inappropriate use.

evaluate the sensitivity of our findings with respect to model classes and use cases.

Table 1.1. Metrics used to evaluate performance, data use, and consent of personalized models and systems. We report performance on a held-out test sample. We assume that individuals report group membership to static models, do not report group membership to imputed models, and only report to participatory systems when informed that it would lead to a strictly positive gain, as computed on the validation set in the training sample.

Metric	Definition	Description
Overall Performance	$\sum_{\boldsymbol{g}\in\mathscr{G}} \frac{n_{\boldsymbol{g}}}{n} \hat{R}_{\boldsymbol{g}}(h_{\boldsymbol{g}})$	Population-level performance of a personalized system/model, computed as a weighted average over all groups
Overall Gain	$\sum_{\boldsymbol{g}\in\mathscr{G}} \frac{n_{\boldsymbol{g}}}{n} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing})$	Population-level gain in performance of a personalized system/model over its generic counterpart
Group Gains	$\min_{\boldsymbol{g}\in\mathscr{G}} / \max_{\boldsymbol{g}\in\mathscr{G}} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing})$	Range of group-level gains of a personalized system/model over its generic counterpart across all groups
Rationality Violations	$\sum_{\boldsymbol{g}\in\mathscr{G}} \mathbb{I}[\text{reject } H_0]$	Number of rationality violations detected using a bootstrap test with 100 resamples at a significance of 10% where $H_0 : \Delta_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\varnothing}) \geq 0$.
Imputation Risk	$\min_{\boldsymbol{g}\in\mathscr{G}} \hat{\Delta}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{g}')$	Worst-case loss in performance due to incorrect imputation. This metric can only be computed for static models.
Options Pruned	$\frac{\lvert\mathscr{R}\rvert - \lvert\mathscr{R}(h)\rvert}{\lvert\mathscr{R}\rvert}$	Proportion of reporting options pruned from a system/model. Here, $\mathscr{R}$ denotes all reporting options and $\mathscr{R}(h)$ denotes those after h is pruned
Data Use	$\sum_{\boldsymbol{g}\in\mathscr{G}} \frac{n_{\boldsymbol{g}}}{n} \frac{\text{requested}(h, \boldsymbol{g})}{\dim(\mathscr{G})}$	Proportion of group attributes requested by h from each group, averaged over all groups in $\mathscr{G}$

1.4.2 Discussion

We show results for logistic regression models and error rate in Table 1.2 and results for other model classes and classification tasks in Appendix A.3. In what follows, we discuss these results.

On Performance

Our results in Table 1.2 show that participatory systems can improve performance across reporting groups. Here, Flat and Seq achieve the best overall performance on 6/6 datasets and improve the gains from personalization for every reporting group on 5/6 datasets. In contrast, traditional models improve overall performance while reducing performance at a group level (see rationality violations on five datasets for 1Hot, mHot). The performance benefits from participatory systems stem from (i) allowing users to opt out of these instances

Table 1.2. Participatory systems and personalized models for all datasets. We summarize metrics in Table 1.1 and present results for other model classes and prediction tasks in Appendix A.3. The best performance across each system is highlighted in green with bold text, and instances of worsenalization are highlighted in red.

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
	Overall Performance	29.1%	29.3%	29.0%	27.9%	28.9%	**24.1%**	24.3%
apnea	Overall Gain	0.1%	-0.1%	0.2%	1.3%	0.3%	**5.1%**	4.9%
$n = 1152, d = 26$	Group Gains	-1.1% – 1.2%	-0.8% – 0.4%	-1.1% – 1.2%	-0.8% – 0.4%	0.0% – 1.2%	0.0% – 13.8%	-0.4% – 13.8%
$\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	1	1	1	1	0	0	0
$\lvert\mathscr{G}\rvert = 6$ groups	Imputation Risk	-4.9%	-5.2%					
Ustun et al. [2016]	Options Pruned	0/6	0/6	0/12	0/12	4/7	5/12	6/12
	Data Use	100.0%	100.0%	0.0%	0.0%	33.3%	83.3%	58.3%
	Overall Performance	21.4%	21.5%	21.6%	22.1%	21.6%	**10.2%**	**10.2%**
cardio_eicu	Overall Gain	0.4%	0.3%	0.3%	-0.2%	0.3%	**11.7%**	**11.7%**
$n = 1341, d = 49$	Group Gains	-1.3% – 2.6%	-2.7% – 3.0%	-1.3% – 2.6%	-2.7% – 3.0%	0.0% – 2.6%	3.1% – 20.9%	3.1% – 20.9%
$\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$	Worsenalization	1	1	1	1	0	0	0
$\lvert\mathscr{G}\rvert = 8$ groups	Imputation Risk	-4.6%	-5.4%					
Pollard et al. [2018]	Options Pruned	0/8	0/8	0/27	0/27	6/9	10/27	9/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	83.3%
	Overall Performance	19.4%	19.3%	19.3%	20.1%	19.2%	**15.7%**	**15.7%**
cardio_mimic	Overall Gain	-0.1%	-0.0%	-0.0%	-0.8%	0.1%	**3.5%**	**3.5%**
$n = 5289, d = 49$	Group Gains	-0.9% – 0.4%	-0.9% – 0.5%	-0.9% – 0.4%	-0.9% – 0.5%	0.0% – 0.4%	-1.6% – 9.8%	-1.6% – 9.8%
$\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$	Worsenalization	3	2	3	2	0	1	1
$\lvert\mathscr{G}\rvert = 8$ groups	Imputation Risk	-1.1%	-1.1%					
Johnson et al. [2016]	Options Pruned	0/8	0/8	0/27	0/27	6/9	6/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	91.7%
	Overall Performance	37.0%	36.7%	37.0%	36.9%	37.0%	36.6%	**36.1%**
coloncancer	Overall Gain	0.1%	0.4%	0.1%	0.2%	0.1%	0.5%	**1.0%**
$n = 29211, d = 72$	Group Gains	-0.4% – 0.3%	-0.1% – 1.1%	-0.4% – 0.3%	-0.1% – 1.1%	0.0% – 0.3%	0.0% – 1.7%	0.2% – 1.7%
$\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	1	0	1	0	0	0	0
$\lvert\mathscr{G}\rvert = 6$ groups	Imputation Risk	-1.4%	-0.9%					
Scosyrev et al. [2012]	Options Pruned	0/6	0/6	0/12	0/12	5/7	7/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	50.0%	75.0%
	Overall Performance	19.6%	19.6%	19.9%	19.8%	19.5%	18.9%	**18.9%**
lungcancer	Overall Gain	-0.1%	-0.1%	-0.3%	-0.2%	0.0%	0.6%	**0.6%**
$n = 120641, d = 84$	Group Gains	-0.4% – 0.2%	-0.3% – 0.2%	-0.4% – 0.2%	-0.3% – 0.2%	0.0% – 0.0%	0.0% – 0.9%	0.3% – 0.9%
$\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$	Worsenalization	4	4	4	4	0	0	0
$\lvert\mathscr{G}\rvert = 6$ groups	Imputation Risk	-0.5%	-0.5%					
Scosyrev et al. [2012]	Options Pruned	0/6	0/6	0/12	0/12	6/7	3/12	7/12
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	83.3%	58.3%
	Overall Performance	20.4%	20.7%	20.4%	29.4%	20.4%	**11.1%**	11.1%
saps	Overall Gain	1.3%	1.0%	1.3%	-7.7%	1.3%	**10.6%**	10.6%
$n = 7797, d = 36$	Group Gains	0.0% – 3.6%	0.0% – 2.7%	0.0% – 3.6%	0.0% – 2.7%	0.0% – 3.6%	4.3% – 17.2%	4.3% – 17.2%
$\mathscr{G} = \{\texttt{HIV}, \texttt{age}\}$	Worsenalization	0	0	0	0	0	0	0
$\lvert\mathscr{G}\rvert = 4$ groups	Imputation Risk	0.0%	-2.4%					
Allyn et al. [2016]	Options Pruned	0/4	0/4	0/9	0/9	1/5	1/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	75.0%	100.0%	75.0%

of "worsenalization" and (ii) assigning personalized predictions with multiple models. Using Table 1.2, we can measure the impact of (i) by comparing the performance of Minimal vs. 1Hot, and the impact of (ii) by comparing the performance of Minimal to Flat (or Seq). For example, on apnea, 1Hot exhibits a significant rationality violation for group [30_to_60,male], meaning they would have been better off with a generic model. By comparing the performance of 1Hot to Minimal, we see that allowing users to opt out of worsenalization reduces test error from 29.1% to 28.9%. By comparing the performance on Minimal to Flat and Seq, we see that using multiple models can further reduce test error from 28.9% to 24.1%.

On Informed Consent

Our results show how Flat and Seq systems can inform consent by allowing users to report a subset of group attributes (e.g., by including reporting options such as $[\texttt{30+}, \varnothing]$ or $[\varnothing, \texttt{HIV+}]$). Although both Flat and Seq systems allow for partial personalization, their capacity to inform consent differs. In a flat system, users may inaccurately gauge the marginal benefit of reporting an attribute by comparing the gains between reporting options. For example, in Fig. 1.4, users who are HIV positive would see a gain of 3.7% for reporting $[\varnothing, \texttt{HIV+}]$, and 16.7% for reporting $[\texttt{30+}, \texttt{HIV+}]$ and may mistakenly conclude that the gain of reporting `age` is $16.7\% - 3.7\% = 13.0\%$. This estimate incorrectly presumes that the gains of 3.7% were distributed equally across age groups. Sequential systems directly inform users of the gains for partial reporting. In the sequential system, group $[\texttt{30+}, \texttt{HIV+}]$ is informed that they would see a marginal gain of 21.5% for reporting `age`, while group $[\texttt{<30}, \texttt{HIV+}]$ is informed they would see a marginal gain of reporting `age` of 0.0%.

On Data Minimization

Our results show that participatory systems perform better across all groups while requesting less personal data on 6/6 datasets. For example, on `cardio_eicu`, Seq reduces error by 11.3% compared to 1Hot while requesting, on average, 83.3% of the data needed by 1Hot. In general, participatory systems can limit data use where personalization does not improve performance, e.g., on `lungcancer`. Even as attributes like `sex` or `age` may be readily reported by patients for any performance benefit, limiting data use is valuable when there is a tangible cost associated with data collection – e.g., when models make use of rating scale for a mental disorder that must be administered by a clinician [Sharp, 2015]. The potential for data minimization varies substantially across prediction tasks. On `apnea`, for example, we can prune six reporting options when building a Seq for decision making (which optimizes error) but four options for Seq for ranking (which optimizes AUC; see Appendix A.3.1). Overall, participatory systems satisfy "global data minimization" as proposed in Biega et al. [2020], in that they minimize the

amount of per-user data requested while achieving the quality of a system with access to the full data on average.

On the Benefits of a Model-Agnostic Approach

Our findings highlight some of the benefits of a model-agnostic approach, in which we can draw on a rich set of models to achieve better performance while mitigating harm. The resulting system can balance training costs with performance benefits. We can also ensure generalization across reporting groups – e.g., by including a generic model fit from a complex model class and personalized models fit from a simpler model class. As expected, fitting for a complex model class can lead to considerable changes in overall accuracy – e.g., we can reduce overall test error for a personalized model from 20.4% to 14.1% on `saps` by fitting a random forest rather than a logistic regression model (see Appendix A.3). However, a gain in overall performance does not always translate to gains at the group level. On `saps`, for example, using a random forest also introduces a rationality violation for one group.

On the Pitfalls of Imputation

One of the simplest approaches to allow individuals to opt out of personalization is to pair a personalized model with an imputation technique. Although this approach can facilitate consent, it may violate the requirements in 1.2. Consider a personalized model that exhibits "worsenalization" in Fig. 1.1. Even if one could correctly impute the group membership for every person, individuals may receive more accurate predictions from a generic model h_0. In practice, imputation is imperfect – as individuals who opt out of reporting their group membership to a personalized model may be assigned "worse" predictions because they are imputed the group membership of a different group. In such cases, opting out may be beneficial, making it difficult for model developers to promote participation while informing consent. Our results highlight the prevalence of these effects in practice. For example, on `cardio_eicu` the estimated "risk of imputation" is -4.6% , indicating that every intersectional group can experience an increase of 4.6% in the error rate as a result of incorrect imputation. The results for KNN-1Hot show that this

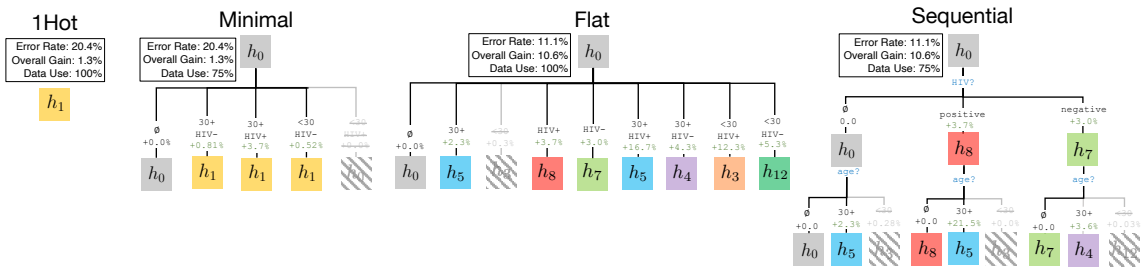


Figure 1.4. Participatory systems for the `saps` dataset. These models predict ICU mortality for groups defined by $\mathscr{G} =$ `HIV` $\times$ `age` $=$ `[+,-]` $\times$ `[<30, 30+]` using logistic regression component models. Here, h_0 is a generic model, h_1 is a 1Hot model fit with a one-hot encoding of $\mathscr{G}$, and $h_2 \cdots h_m$ are 1Hot and mHot models fit for each reporting group. We show the gains of each reporting option above each box and highlight pruned options in grey. For example, in Seq, the group `(HIV+, 30+)` sees an estimated 21.5% error reduction after reporting `HIV` if they report `age`. In contrast, the group `(HIV+, <30)` sees no gain from reporting `age` in addition to `HIV` status, so this option is pruned.

potential harm can be realized in practice using KNN-imputation, as we find that the imputed system leads to rationality violations on 5/6 datasets.

1.5 Concluding Remarks

We introduced a new family of classification models that allow individuals to report personal data at prediction time. Our work focuses on personalization with group attributes; our approach could be used to facilitate and inform consent in a broader class of prediction tasks. In such cases, the key requirement for building a participatory system is that we can reliably estimate the gains of personalization for each person who reports personal data.

Our results show that participatory systems can inform consent while improving performance and reducing data use across groups. Reaping these benefits in practice will hinge on the ability to effectively inform decision subjects on the impact of their reporting decisions. Anderson and Agarwal [2011]. Even as there may be good "default practices" for what kind of information we should show decision subjects, practitioners should tailor this information to the application and target audience [Edwards et al., 2013].

One common concern in using a participatory system arises when practitioners wish to

collect data from a model in deployment to improve its performance in the future. In practice, a participatory system can thwart data collection in such settings by allowing individuals to opt out. In such cases, we would note that this issue should be resolved in a way that is aligned with the principle of *purpose specification* OECD [2013]. If the goal of data collection is to improve a model, then individuals could always be asked to report information voluntarily for this purpose. If the goal of data collection is to personalize predictions, then individuals should be able to opt out, especially when it may lead to worse performance.

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Neural Information Processing Systems 36 (NeurIPS 2023). Hailey Joren; Chirag Nagpal; Katherine A. Heller; and Berk Ustun. The dissertation author was the primary investigator and author of this paper. We thank the following individuals for helpful discussions: Taylor Joren, Sanmi Koyejo, Charlie Marx, Julian McAuley, and Nisarg Shah. This work was supported by funding from the National Science Foundation IIS 2040880, the NIH Bridge2AI Center Grant U54HG012510, and an Amazon Research Award.

Chapter 2

Classification with Conceptual Safeguards

2.1 Introduction

One of the most promising applications of machine learning is to automate routine tasks that a human can perform. We can now train deep learning models to perform such tasks across applications – be it to identify a bird in an image [Tabak et al., 2019], detect toxicity in text [Kumar et al., 2021], or diagnose pneumonia in a chest x-ray [Yan et al., 2023]. Even as these models may outperform human experts [He et al., 2015, Zhou et al., 2021, Chong et al., 2020], their performance falls short of the levels we need to reap the benefits of full automation. A bird identification model that is 80% accurate may not work reliably enough to be used in the field by conservationists. A pneumonia detection model with 95% accuracy may still not be sufficient to eliminate the need for human oversight in a hospital setting.

One strategy to reap some of the benefits of automation in such tasks is through *abstention*. Given any model that is insufficiently accurate, we can measure the uncertainty in its predictions and improve its accuracy by abstaining from predictions that are too uncertain. In this way, we can improve accuracy by sacrificing *coverage* – i.e., the proportion of instances where a model assigns a prediction. One of the common barriers to abstention in general applications is how to handle instances where a model abstains. In applications where we wish to automate a routine task, we would pass these to the human expert who would have made the decision in the first place. Thus, abstention represents a way to reap benefits from *partial automation*.

In this paper, we present a modeling paradigm to ensure safety in such applications that we call a *conceptual safeguard* (see Fig. 2.1). A conceptual safeguard is an abstention mechanism for models that predict an outcome by first predicting a set of concepts. Given such a model, a conceptual safeguard operates as a verification layer – estimating the uncertainty in each prediction and abstaining when it exceeds a threshold needed to ensure a minimal level of accuracy. Unlike a traditional selective classifier, a conceptual safeguard provides a way to improve coverage – by allowing experts to confirm the presence of certain concepts on instances on which we abstain.

Although conceptual safeguards are designed to be a simple component that can be implemented off the shelf, designing methods to learn them is challenging. On the one hand, concepts introduce a degree of uncertainty that we must account for at test time. In the best case, we may have to abstain too much to achieve a desired level of accuracy. In the worst case, we may fail to hit the mark. On the other hand, we must build systems that are *designed for confirmation* – i.e., where we can reasonably expect that confirming concepts will improve coverage and where we can rank the concepts in a way that improves coverage. Our work seeks to address these challenges so that we can reap the benefits of these models.

Contributions

Our main contributions include:

1. We introduce a general-purpose approach for safe automation in classification tasks with concept annotation. Our approach can be applied using off-the-shelf supervised learning techniques.

2. We develop a technique to account for concept uncertainty in concept bottleneck models. Our technique can use native estimates from the components of a concept bottleneck to return reliable estimates of label uncertainty, improving the accuracy-coverage trade-off in selective classification.

3. We propose a confirmation policy that can improve coverage. Our policy flags concepts in instances on which a model abstains, prioritizing high-value concepts that could resolve abstention. This strategy can be readily customized to conform to a budget and applied offline.

4. We benchmark conceptual safeguards on classification datasets with concept labels. Our results show that safeguards can improve accuracy and coverage through uncertainty propagation and concept confirmation.

Related Work

Selective Classification

Our work is related to a large body of work on machine learning with a reject option [see e.g., Hendrickx et al., 2021, for a recent survey], and more specifically, methods for selective classification [Chow, 1957, 1970, Fumera and Roli, 2000, Franc et al., 2023]. Our goal is to learn a selective classifier that assigns as many predictions as possible while adhering to a minimal level of accuracy [i.e., the bounded abstention model of Franc et al., 2023]. We build conceptual safeguards for this task through a *post-hoc* approach [see e.g., Geifman and El-Yaniv, 2017] – in which we are given a model, and build a verification layer that estimates the confidence of each prediction and abstains on predictions where a model is insufficiently confident. The key challenge in our approach is that we require reliable estimates of uncertainty to abstain effectively [Charoenphakdee et al., 2021, Fisch et al., 2022], which we address by propagating the uncertainty in concept predictions to the uncertainty in the final outcome.

Deep Learning with Concepts

Our work is related to a stream of work on deep learning with concept annotations. A large body of work uses these annotations to promote interpretability – either by explaining the predictions of DNN in terms of concepts that a human can understand [Kim et al., 2018, Chen et al., 2020], or by building *concept bottleneck models* – i.e., a model that predicts an outcome of interest by predicting concepts – i.e., [Koh et al., 2020]. One of the key motivations

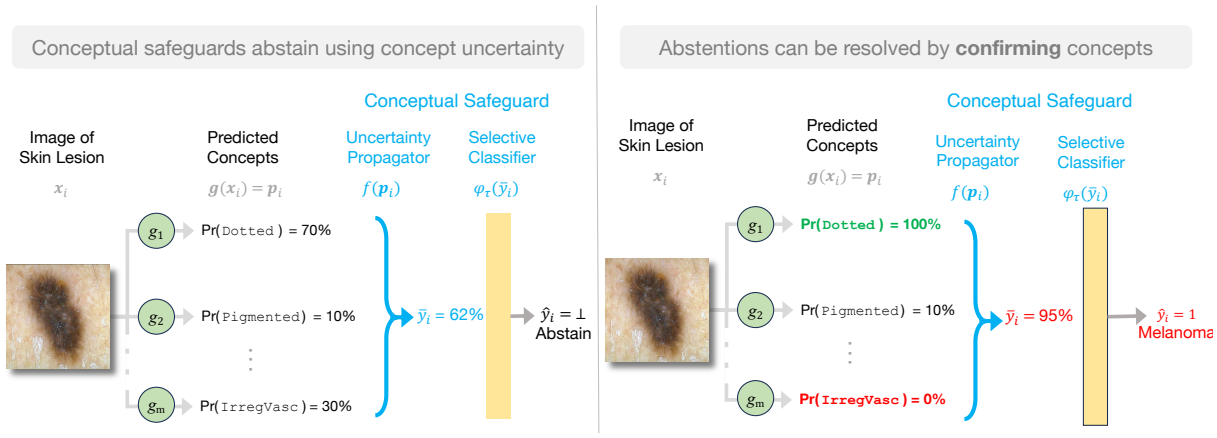


Figure 2.1. Conceptual safeguard to detect melanoma from an image of a skin lesion. We consider a model that estimates the probabilities of m concepts: `Dotted`, `Pigmented` ... `IrregularVasc`. Given these probabilities, a conceptual safeguard will decide whether to output a prediction $\hat{y} \in \{\texttt{Melanoma}, \texttt{NoMelanoma}\}$ or to abstain $\hat{y} = \perp$. The safeguard improves accuracy by abstaining on images that would receive a low confidence prediction, and measures confidence in a way that accounts for the uncertainty in concept predictions through uncertainty propagation. On the left, we show an image where a safeguard abstains because its confidence fails to meet the threshold to ensure high accuracy $\Pr(\texttt{Melanoma}) = 62\% \leq 90\%$. On the right, we show that a human expert can resolve the abstention by confirming the presence of concepts `Dotted` and `IrregVasc` in the image.

for concept bottleneck models is the potential for humans to intervene at test time [Koh et al., 2020] – e.g., to improve performance by correcting a concept prediction that was incorrectly detected. Recent work shows that many architectures that perform well cannot readily support interventions [see e.g., Havasi et al., 2022, Margeloiu et al., 2021].[1] Likewise, interventions may not be practical in applications where we wish to automate a routine task. In such cases, we need a mechanism to flag predictions for human review to prevent human experts from having to check each prediction. Our work highlights several avenues to avoid these limitations. In particular, we work with an independent architecture that is amenable to interventions, consider a restricted class of interventions, and present an approach that does not require constant human supervision.

One overarching challenge in building concept bottleneck models is the need for concept

[1]For example, if we build a sequential architecture – wherein we train a front-end model to predict the label from *predicted concepts* – then interventions may reduce accuracy as the front-end may rely on an incorrect concept prediction to output accurate label predictions.

annotations. In effect, very few datasets include concept annotations, and those that do are often incomplete (e.g., an example may be missing some or all concept labels). In practice, the lack of concept labels can limit the applicability and the performance of concept bottleneck models – and has motivated a recent stream of work on machine-driven concept annotation [see e.g., Oikarinen et al., 2023] and drawing on alternate sources of information [Yuksekgonul et al., 2022, Yeh et al., 2020]. Our work outlines an alternative approach to overcome this barrier to adoption: rather than building a new model that is sufficiently accurate, use it as much as possible by allowing it to abstain from prediction.

2.2 Framework

We consider a classification task to predict a label from a complex feature space. We start with a dataset of n i.i.d. training examples $\{(\mathbf{x}_i, \mathbf{c}_i, y_i)\}_{i=1}^n$, where example i consists of:

- a vector of *features* $\mathbf{x}_i \in \mathscr{X} \subseteq \mathbb{R}^d$ – e.g., $\mathbf{x}_i$ pixels in image i;
- a vector of k *concepts* $\mathbf{c}_i \in \mathscr{C} = \{0,1\}^m$ – e.g., $c_{i,k} = 1$ if x-ray i contains a bone spur;
- a *label* $y_i \in \mathscr{Y} = \{0,1\}$ – e.g., $y_i = 1$ if patient i has arthritis.

Objective

We use the dataset to build a selective classification model $h : \mathscr{X} \to \mathscr{Y} \cup \{\perp\}$. Given a feature vector $\mathbf{x}_i$, we denote the output of the model as $\hat{y}_i := h(\mathbf{x}_i)$ where $\hat{y}_i = \perp$ denotes that the model *abstains from prediction* for $\boldsymbol{x}_i$.

In the context of an automation task, we would like our model to assign as many predictions as possible while adhering to a target accuracy to ensure safety at test time. In this setup, abstention reflects a viable path to meet this constraint – by allowing the model to abstain on instances where it would otherwise assign an incorrect prediction. Given an *target accuracy*

$\alpha \in (0,1)$, we express these requirements as an optimization problem:

$$\begin{aligned} \max_{h} \quad & \text{Coverage}(h) \\ \text{s.t.} \quad & \text{Accuracy}(h) \geq \alpha, \end{aligned}$$

where:

- $\text{Coverage}(h) := \Pr(\hat{y} \neq \perp)$ is the *coverage* of the model h – i.e., proportion of instances where h outputs a prediction.
- $\text{Accuracy}(h) := \Pr(y = \hat{y} \mid \hat{y} \neq \perp)$ is the *selective accuracy* of the model h – i.e., the accuracy of h over instances on which it outputs a prediction

System Components

We consider a selective classifier with the components shown in Fig. 2.1. Given the dataset, we will first train the basic components of an independent concept bottleneck model:

- A *concept detector* $g : \mathscr{X} \to [0,1]^m$. Given features $\boldsymbol{x}_i$, the concept detector returns as output a vector of m probabilities $\mathbf{q}_i = [q_{i,1}, \ldots, q_{i,m}] \in [0,1]^m$ where each $q_{i,k} := \Pr(c_{i,k} = 1 \mid \boldsymbol{x}_i)$ captures the probability that concept k is present in instance i.
- A *front-end model* $f : \mathscr{C} \to \mathscr{Y}$, which takes as input a vector of *hard* concepts $\mathbf{c}_i$ and returns as output the outcome probability $\bar{y}_i := f(\mathbf{c}_i)$.

In what follows, we treat these models as fixed components that are given and assume that they satisfy two requirements: (i) *independent training*, meaning that we train the concept detector and the front-end model using the datasets $\{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^n$ and $\{(\mathbf{c}_i, y_i)\}_{i=1}^n$, respectively; (ii) *calibration*, i.e., that the concept detector and front-end model and return calibrated probability estimates of their respective outcomes. As we will discuss shortly, independent training will be essential for confirmation, while calibration will be essential for abstention.

In practice, these are mild requirements that we can satisfy using off-the-shelf methods for supervised learning. Given a dataset, for example, we can fit the concept detectors using

a standard deep learning algorithm, and fit the front-end model using logistic regression. In settings where the dataset is missing concept labels for certain examples, we can train a separate detector for each concept to maximize the amount of data for each concept detector. In settings where we have an embedding model [Yuksekgonul et al., 2022] or an existing end-to-end deep learning model, we can dramatically reduce the computation by training the concept detectors via fine-tuning. In all cases, we can calibrate each model by applying a post-hoc calibration technique, e.g., Platt scaling [Platt et al., 1999].

Conceptual Safeguards

Conceptual safeguards operate as a confidence-based selective classifier – abstaining on points where they cannot assign a sufficiently confident prediction. We control this behavior through an internal component called the *selection gate* $\varphi_\tau : [0,1] \to \mathscr{Y} \cup \{\perp\}$, which is parameterized with a *confidence threshold* $\tau \in [0,1]$. Given a threshold τ, the gate takes as input a soft label $\bar{y}_i \in [0,1]$ and returns:

$$\hat{y}_i = \varphi_\tau(\bar{y}_i) = \begin{cases} 0 & \text{if} \quad \bar{y}_i \in [0, \tau) \\ \perp & \text{if} \quad \bar{y}_i \in [\tau, 1-\tau] \\ 1 & \text{if} \quad \bar{y}_i \in (1-\tau, 1] \end{cases}$$

As shown in Fig. 2.2, we can set τ to choose an operating point for the model on the accuracy-coverage curve – e.g., to meet a desired target accuracy rate by sacrificing coverage. In settings where our front-end model returns calibrated probability predictions, we can use them to set the threshold τ.

Definition 1. A probabilistic prediction $\bar{y} \in [0,1]$ for a binary label $y \in \{0,1\}$ is *calibrated* if $\Pr(y = 1 \mid \bar{y} = t) = t$ for all values $t \in [0,1]$.

Proposition 1. *Suppose that $\bar{y}$ is a calibrated probability prediction for y. Then any selective classifier $\varphi_\tau(\bar{y})$ that abstains when $\bar{y}$ has confidence below $1-\tau$ achieves accuracy at least $1-\tau$.*

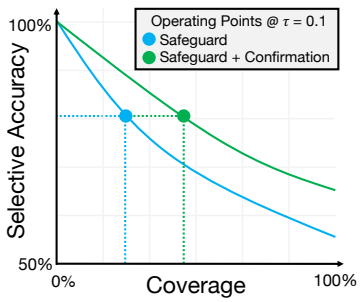


Figure 2.2. We evaluate the performance of classification models that can abstain using an accuracy-coverage curve [Franc et al., 2023]. Given a model that outputs a probability prediction for each point, a conceptual safeguard flags points on which a model abstains based on a confidence threshold $\tau \in [0, 0.5]$ – where setting $\tau = 0$ leads to 100% coverage and setting $\tau = 0.5$ leads to 0% coverage.

Proposition 1 is a standard result ensuring high accuracy for selective classifiers that output calibrated probabilistic prediction (see Appendix B.1 for a proof). In cases where the front-end model is not calibrated, the result will hold only given the degree of calibration. Thus, the reliability of this approach hinges on the reliability of our confidence estimates. An alternative approach for such cases is to treat the output of the front-end model f as a *confidence score* and tune τ using a calibration dataset [see, e.g., the SGR algorithm of Geifman and El-Yaniv, 2017].

Confirmation

We let users *confirm* the presence of concepts on instances where a model abstains. In a pneumonia detection task, for example, we can ask a radiologist to confirm the presence of concept k in x-ray i. We refer to this procedure as confirmation rather than intervention to distinguish it from other ways where a human expert would alter the output from a concept detector.[2]

We consider tasks where each concept is *human-verifiable* – i.e., that it can be detected correctly by the human experts that we would query to confirm [c.f., concepts where a human may be uncertain as in Collins et al., 2023]. In such tasks, confirming a concept will replace its

[2]For example, intervention may refer to "correction" in which a human replaces a hard concept prediction with its correct value.

probability $q_{i,k}$ with its ground-truth value $c_{i,k} \in \{0,1\}$. We cast confirmation as a *policy* function $\psi_S : [0,1]^m \to [0,1]^m$ where $S \subseteq [m]$ denotes a subset of concepts to confirm. The function takes as input a vector of raw concept predictions and returns a vector of partially confirmed concept predictions: $p_i = [p_{i,1}, \ldots, p_{i,m}] \in [0,1]^m$ where:

$$p_{i,k} := \begin{cases} c_{i,k} & \text{if } k \in S \\ q_{i,k} & \text{if } k \notin S \end{cases}$$

2.3 Methodology

In this section, we describe how to build the internal components of a conceptual safeguard. We first discuss how to output reliable predicted probabilities given uncertain inputs at test time. We then introduce a technique to flag promising examples that can be confirmed to improve coverage.

2.3.1 Uncertainty Propagation

It is important that the concept detectors are probabilistic, so that we can prioritize confirming concepts that have high uncertainty. However, this creates an issue where the concept detectors output probabilities, but the front-end model requires hard concepts as inputs.

Here, we describe a simple strategy wherein we use *uncertainty propagation* to allow the front-end model to accept probabilities rather than hard concepts as input. We make two assumptions about the underlying data distribution to motivate our approach.

Assumption 1. *The label y and features* $\boldsymbol{x}$ *are conditionally independent given the concepts* $\boldsymbol{c}$.

Assumption 2. *The concepts* $\{c_1, \ldots, c_m\}$ *are conditionally independent given the features* $\boldsymbol{x}$.

We can use these assumptions to write the conditional label distribution $p(y \mid \boldsymbol{x})$ in terms

of quantities that we can readily obtain from a concept detector and front-end model:

$$p(y \mid \boldsymbol{x}) = \sum_{\mathbf{c} \in \{0,1\}^m} p(y \mid \boldsymbol{c}, \boldsymbol{x})\, p(\mathbf{c} \mid \mathbf{x}) \tag{2.1}$$

$$= \sum_{\mathbf{c} \in \{0,1\}^m} p(y \mid \mathbf{c})\, p(\boldsymbol{c} \mid \mathbf{x}) \qquad \text{(Assumption 1)}$$

$$= \sum_{\mathbf{c} \in \{0,1\}^m} \underbrace{p(y \mid \boldsymbol{c})}_{\substack{\text{output from} \\ f(\mathbf{c})}} \prod_{k \in [m]} \underbrace{p(c_k \mid \boldsymbol{x})}_{\substack{\text{output from} \\ g_k(\mathbf{x})}} \qquad \text{(Assumption 2)}$$

We use this decomposition to propagate uncertainty from the inputs of the front-end model to its output. Specifically, we compute the expected prediction of the front-end model on all possible realizations of hard concepts and weigh each realization in terms of the probabilities from concept detectors. In practice, we replace each quantity in Eq. (2.1) with an estimate computed by the concept detectors and front-end model. Given predicted probabilities from concept detectors $p_i = (p_{i,1}, \ldots, p_{i,k})$, we compute the estimate of $p(y \mid \boldsymbol{x})$ as:

$$f(p_i) := \sum_{\boldsymbol{c} \in \{0,1\}^m} f(\boldsymbol{c}) \prod_{k \in [m]} p_{i,k}^{c_k} (1 - p_{i,k})^{1 - c_k} \tag{2.2}$$

$$\underbrace{f(p_i)}_{\substack{\text{output using} \\ \text{uncertain} \\ \text{concepts}}} := \sum_{\mathbf{c} \in \{0,1\}^m} \underbrace{p(y \mid \boldsymbol{c})}_{\substack{\text{output using} \\ \text{hard concepts}}} \prod_{k \in [m]} \underbrace{p(c_k \mid \boldsymbol{x})}_{\substack{\text{output from} \\ \text{concept} \\ \text{detectors}}} \tag{2.3}$$

Here, we abuse notation slightly and use $f(p_i)$ to denote the front-end model applied to uncertain concepts, and $f(\mathbf{c})$ to denote the front-end model applied to hard concepts. This requires 2^m calls to the front-end model, which is negligible in practice as most front-end models are trained with a limited number of concepts. In settings where m is large, or computation is prohibitive,

we can use a sample of concept vectors to construct a Monte Carlo estimate.

2.3.2 Confirmation

Selective classification guarantees higher accuracy at the cost of potentially lower coverage. In what follows, we describe a strategy that can mitigate the loss in confirmation by human confirmation – i.e., manually spotting concepts among instances on which we abstain. In principle, confirmation will always lead to an improvement in coverage. In practice, human confirmation is labor-intensive – and may require expertise – so we want to develop techniques that can account for this by flagging promising examples that are responsive to confirmation costs.

In Algorithm 2, we present a routine to flag concepts for a human expert to review among instances on which a model abstains. The routine iterates over a batch of n points on which the model has abstained and identifies salient concepts that can be confirmed by a human expert to avoid abstention.

Algorithm 2 computes the gain associated with confirming each concept on each instance and then returns the concepts with the maximum gain while adhering to a user-specified *confirmation budget* $B > 0$. We associate the cost of confirming each concept k with a cost $\gamma_k > 0$, which can be set to control the time or expertise that is associated with each confirmation. The routine selects concepts for review based on the expectation of the gain in certainty. We measure the gain in certainty in terms of the variance of the prediction after confirmation:

$$\begin{aligned} \text{Gain}(\mathbf{q}, k) &= \text{Var}_{p_k \sim \text{Bern}(q_k)} \left[f(\psi_{\{k\}}(\mathbf{q})) \right] \\ &= q_k(1-q_k) \left(f(\mathbf{q}[q_k \leftarrow 1]) - f(\mathbf{q}[q_k \leftarrow 0]) \right)^2 \end{aligned} \tag{2.4}$$

Algorithm 2. Greedy Concept Selection

Require: $\{i \in [n] \mid \varphi_\tau(f(\mathbf{q}_i)) = \perp\}$ instances on which a model abstained
Require: $\gamma_1, \ldots, \gamma_m > 0$, cost to confirm each concept
Require: $B > 0$, confirmation budget
$S_1, \ldots, S_n \leftarrow \{\}$ *concepts to confirm for each instance*
repeat
  $i^*, k^* \leftarrow \arg\max_{i,k} \text{Gain}(\mathbf{q}_i, k)$ s.t. $k \notin S_i$ *select best remaining concept*
  $S_{i^*} \leftarrow S_{i^*} \cup \{k^*\}$
  $B \leftarrow B - \gamma_{k^*}$
until $B < 0$ or $S_i = [m]$ for all $i \in [n]$
Output $S_1, \ldots, S_n$, concepts to confirm for each abstained instance

$$\underbrace{\text{Gain}(\mathbf{q}, k)}_{\substack{\text{gain from} \\ \text{confirming} \\ \text{concept}}} := q_k(1 - q_k)\underbrace{f(\mathbf{q}[q_k \leftarrow 1])}_{\substack{\text{outcome if} \\ \text{concept present}}} - \underbrace{f(\mathbf{q}[q_k \leftarrow 0])}_{\substack{\text{outcome if} \\ \text{concept absent}}} \tag{2.5}$$

The gain measure in (2.4) captures the sensitivity of predictions from the front-end model by confirming concept k. In particular, we seek to identify concepts that – if confirmed – would induce a large change in the output of the front-end and thus resolve abstentions. Given that we do not know the underlying value of concept k prior to confirmation, we treat q_k as a random variable that will be set to $\mathbf{q}[q_k \leftarrow 1]$ with probability q_k and $\mathbf{q}[q_k \leftarrow 0]$ with probability $1 - q_k$. Here, $\mathbf{q}[q_k \leftarrow 1]$ refer to the vector $\mathbf{q}$, except with q_k replaced with 1.

2.4 Experiments

We present experiments where we benchmark conceptual safeguards on a collection of real-world classification datasets. Our goal is to evaluate their accuracy and coverage trade-offs, and to study the effect of uncertainty propagation and confirmation through ablation studies. We include details on our setup and results in Appendix B.2, and provide code to reproduce our results on GitHub.

2.4.1 Setup

We consider six classification datasets with concept annotations:

- The `melanoma` and `skincancer` datasets are image classification tasks to diagnose melanoma and skin cancer derived from the Derm7pt dataset [Kawahara et al., 2018].
- The `warbler` and `flycatcher` datasets are image classification tasks derived from the CalTech-UCSD Birds dataset [Wah et al., 2011] to identify different species of birds.
- The `noisyconcepts25` and `noisyconcepts75` datasets are synthetic classification tasks designed to control the noise in concepts (see Appendix B.2.1).

We process each dataset to binarize categorical concepts (e.g., **WingColor** to **WingColorRed**). We split each dataset into a training sample (80%, used to build a selective classification model) and a test sample (20%, used to evaluate coverage and selective accuracy in deployment).

Models

We train a selective classification model using one of the following methods:

- X→Y MLP: A multilayer perceptron trained on top of the penultimate layer of the embedding model. This baseline represents an end-to-end deep learning model that directly predicts the output without concepts.
- Baseline: An independent concept bottleneck model built from concept detectors $g_1, \ldots, g_m$ and a front-end model f trained to predict the true concepts.
- CS: Conceptual safeguard built from the same concept detectors and front-end as the baseline. This model propagates the uncertainty from the concept detectors $g_1, \ldots, g_m$ to the front-end model f as described in Section 2.3.

We build Baseline and CS models using the same front-end model f and concept detectors $g_1, \ldots, g_m$. We train the front-end model f using logistic regression, and the concept detectors using embeddings from a pre-trained model [i.e., InceptionV3, Szegedy et al., 2016] (for all datasets other than `noisyconcepts`).

Evaluation

We report the performance of each model through an *accuracy-coverage curve* as in Fig. 2.2, which plots its coverage and selective accuracy on the test sample across thresholds. We evaluate the impact of confirmation in concept based models – i.e., Baseline and CS – in terms of the following policies:

- ImpactConf, which flags concepts to review using Algorithm 2;
- RandomConf, which flags a random subset of concepts to review.

We control the number of examples to confirm by setting a *confirmation budget*, and plot accuracy-coverage curves for confirmation budgets of 0/10/20/50% to show how performance changes under no/low/medium/high human levels of human intervention, respectively.

2.4.2 Results

We present the accuracy coverage curves for all methods and all datasets in Table 2.1. Given these curves, we can evaluate the gains to uncertainty propagation by comparing Baseline to CS, and the gains from confirmation by comparing Baseline + RandomConf to CS + ImpactConf).

Overall, our results show that our methods outperform their respective baselines in terms of coverage and selective accuracy. In general, we find that the gains vary across datasets and confirmation budgets. Given a desired target accuracy, for example, we achieve higher coverage on `warbler` and `flycatcher` rather than `melanoma` and `skincancer`. Such differences arise due to differences in the quality of concept annotations and their relevance for the prediction task at hand.

On Uncertainty Propagation

Our results highlight how uncertainty propagation can lead to improvements in a safeguard. On the one hand, we find that uncertainty propagation improves the accuracy coverage trade-off. On the `skincancer` dataset, for example, we see a major improvement in the accuracy coverage curves between a model that propagates uncertainty (CS + RandomConf, blue) and a

Table 2.1. Accuracy coverage curves for all methods on all datasets. We include additional results in Appendix B.2 for multiclass tasks. Note that Baseline (black) and Baseline + RandomConf (green) produce identical results without confirmation. Likewise, CS + RandomConf (blue) and CS + ImpactConf (purple) are also equivalent under the same condition.

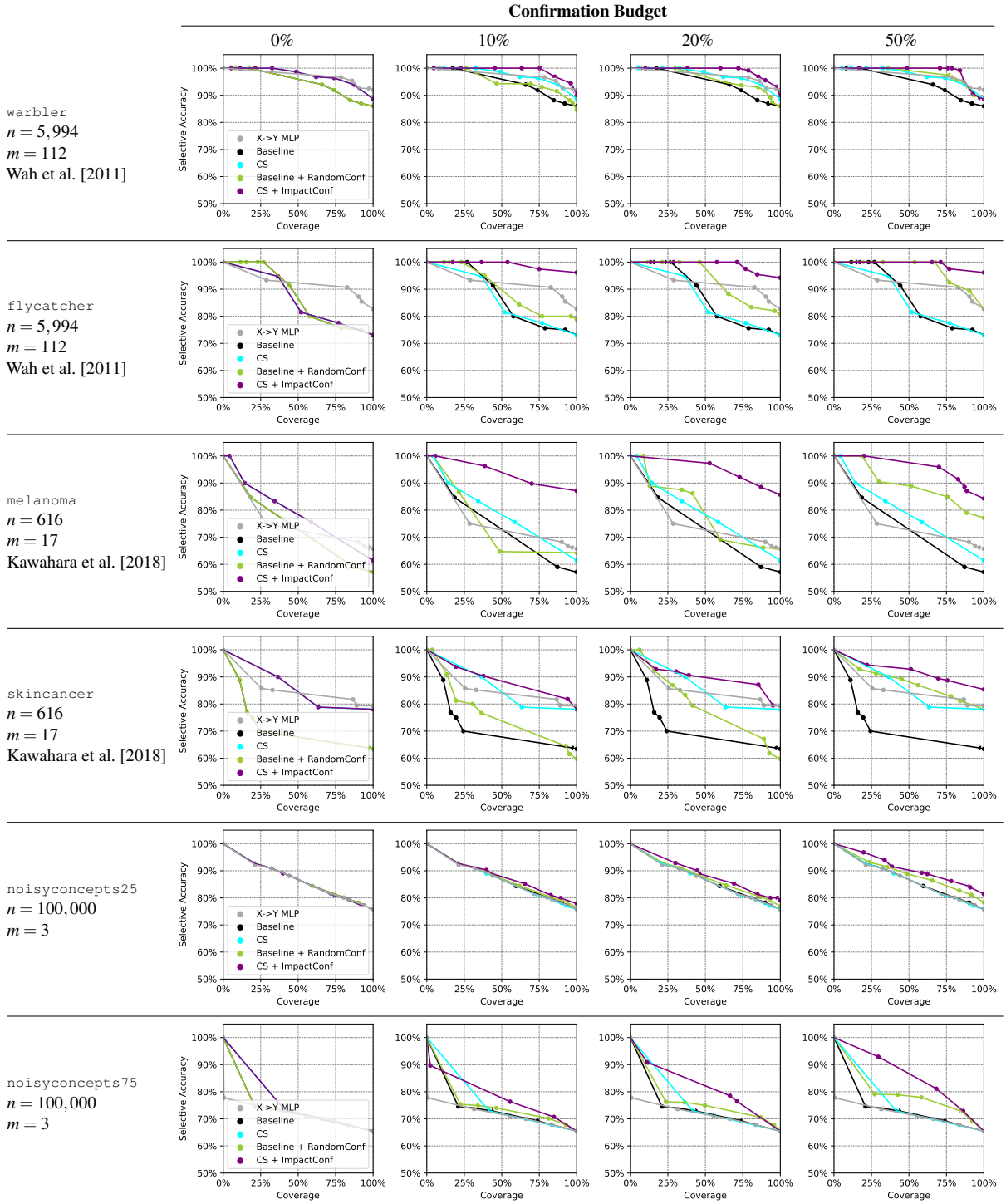


Table 2.2. Coverage at specific thresholds τ for a confirmation budget of 20%. We present results for other datasets and confirmation budgets in Appendix B.2.2.

Dataset	Prediction Thresholds	X->Y MLP	Baseline	CS	Baseline + RandomConf	CS + ImpactConf
warbler	$\tau = 0.05$	86.0%	17.3%	74.0%	30.0%	90.00%
$n = 5,994$	$\tau = 0.1$	100.00%	74.0%	87.3%	89.3%	100.00%
$m = 112$	$\tau = 0.15$	100.00%	100.00%	100.00%	100.00%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	100.00%	100.00%	100.00%	100.00%
flycatcher	$\tau = 0.05$	0.0%	26.9%	0.0%	46.2%	84.62%
$n = 5,994$	$\tau = 0.1$	82.7%	44.2%	36.5%	46.2%	100.00%
$m = 112$	$\tau = 0.15$	92.3%	44.2%	36.5%	65.4%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	57.7%	51.9%	100.00%	100.00%
melanoma	$\tau = 0.05$	0.0%	0.0%	4.3%	8.6%	52.86%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	14.3%	8.6%	72.86%
$m = 17$	$\tau = 0.15$	0.0%	0.0%	14.3%	41.4%	100.00%
Kawahara et al. [2018]	$\tau = 0.2$	0.0%	18.6%	34.3%	41.4%	100.00%
skincancer	$\tau = 0.05$	0.0%	0.0%	0.0%	6.10%	0.0%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	36.6%	15.9%	39.02%
$m = 17$	$\tau = 0.15$	32.9%	11.0%	36.6%	28.0%	85.37%
Kawahara et al. [2018]	$\tau = 0.2$	86.59%	11.0%	36.6%	36.6%	85.4%
noisyconcepts25	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 100,000$	$\tau = 0.1$	32.3%	32.3%	32.3%	33.9%	44.66%
$m = 3$	$\tau = 0.15$	44.1%	43.6%	43.6%	45.8%	69.16%
	$\tau = 0.2$	80.4%	80.4%	80.4%	86.8%	93.31%
noisyconcepts75	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 100,000$	$\tau = 0.1$	0.0%	0.0%	0.0%	0.0%	11.17%
$m = 3$	$\tau = 0.15$	0.0%	0.0%	0.0%	0.0%	11.17%
	$\tau = 0.2$	0.0%	0.0%	0.0%	0.0%	11.17%

model and a comparable model that does not (Baseline + RandomConf, green).

On the other hand, accounting for uncertainty can improve these trade-offs by producing a more effective confirmation policy. Our results highlight these effects by showing the gains of uncertainty may change under a confirmation budget. On the flycatcher dataset, for example, accounting for uncertainty leads to little difference when we do not confirm examples (i.e., a 0% confirmation budget). In a regime where we allow for confirmation – setting a 20% confirmation budget – we find that accounting for uncertainty can increase coverage, from 46.2% to 63.5%, when the threshold is set at $\tau = 0.05$ (see Section 2.4.1).

On Confirmation

Our results show that confirming concepts improves performance across all datasets. On flycatcher, we find that confirming a random subset of instances 20% improves coverage from 26.9% to 46.2% for a threshold of $\tau = 0.05$ (Baseline $\rightarrow$ Baseline + RandomConf). In a conceptual safeguard, coverage starts from 63.5% due to uncertainty propagation (CS + RandConf), and

improves to 84.6% as a result of our targeted confirmation policy (CS + ImpactConf). The gains of confirmation depend on the underlying task and dataset. For example, the gains may be smaller when the concept detectors perform well enough without confirmation to achieve high accuracy (e.g., for the `warbler` and `noisyconcepts25` datasets).

Our results highlight the value of targeted confirmation policy – i.e., as a technique that can lead to meaningful gains in coverage without compromising safety or requiring real-time human supervision. In practice, these gains only part of the benefits of our approach – as practitioners may be able to specify costs in a way that limits the experts required for confirmation. For example, non-expert users may be able to confirm concepts such as `WingColorRed`, while confirming the final species prediction to `RedFacedCormorant` may require greater expertise.

2.5 Concluding Remarks

Conceptual safeguards reflect a general-purpose approach to promote safety through selective classification. In applications where we wish to automate routine tasks that humans could perform, safeguards allow us to reap some of the benefits of automation through abstention in a way that can promote interpretability and improve coverage. Although our work has primarily focused on binary classification tasks, our machinery can be applied to build conceptual safeguards for multiclass tasks (see Appendix B.2.3), and extended to other supervised prediction problems.

Our approach has a number of overarching limitations that affect concept bottlenecks and selective classifiers. As with all models trained with concept annotations, we expect to incur some loss in performance relative to an end-to-end model when concepts are unlikely to capture all relevant information about the label from its input. In practice, this gap may be large – and we may reap the benefits of automation using a traditional selective classifier. As with other selective classification methods, we expect that abstention may exacerbate disparities in coverage or performance across subpopulations [see e.g., Jones et al., 2021]. In our setting, it may be

difficult to pin down the source of these disparities – as they may arise from concept annotations, model concepts, or interaction effects. Nevertheless, we may be able to mitigate these effects through a suitable confirmation policy.

Acknowledgments

Chapter 2, in full, is a reprint of the material as it appears in The Twelfth International Conference on Learning Representations (ICLR 2024). Hailey Joren; Charles Marx and Berk Ustun. The dissertation author was the primary investigator and author of this paper. We thank Taylor Joren, Mert Yuksekgonul, and Lily Weng for helpful discussions. This work was supported by funding from the National Science Foundation Grants IIS-2040880 and IIS-2313105, the NIH Bridge2AI Center Grant U54HG012510, and an Amazon Research Award.

Chapter 3

Sufficient Context: A New Lens on Retrieval Augmented Generation Systems

3.1 Introduction

Providing Large Language Models (LLMs) with additional context, such as in Retrieval Augmented Generation (RAG) systems, has led to major improvements in LLM factuality and verifiability when adapting to new domains [Lewis et al., 2020]. In the case of open-domain question answering, a retrieval model provides context at inference time in the form of snippets or long-form text [Zhu et al., 2021]. Then, the model synthesizes the query along with this added context to generate the answer. Unfortunately, current RAG-based LLMs exhibit many undesirable traits, such as confidently predicting the incorrect answer with retrieved evidence [Mishra et al., 2024, Niu et al., 2024, Ru et al., 2024], being distracted by unrelated information [Cuconasu et al., 2024, Yoran et al., 2024], and failing to properly extract answers from long text snippets [Hsieh et al., 2024, Liu et al., 2024].

The ideal outcome is for the LLM to output the correct answer if the provided context contains enough information to answer the question when combined with the model's parametric knowledge. Otherwise, the model should abstain from answering and/or ask for more information. One core challenge in achieving this ideal outcome is building models that can use the provided context only when it helps answer the question correctly. Several works have investigated this issue by evaluating models in the presence of irrelevant information in the context (discussed in

Section 3.2). However, "relevant information" can range from directly containing the answer to simply being topically related to the question. Even "golden" or oracle documents in datasets vary in how much information they provide about the query, and whether they directly inform the ground truth answer or not. In other words, while the goal seems to be to understand how LLMs behave when they do or do not have sufficient information to answer the query, prior work fails to address this head-on.

As our first contribution, we put forth a new notion of sufficient context. We divide instances into two categories based on whether the context provides enough information to construct an answer to the query. The sufficient context designation is a function of an input pair consisting of one question and the associated context. Crucially, it does not require a ground truth answer. Figure 1 shows examples and a breakdown of model responses after splitting the data based on sufficient vs. insufficient context. To divide the dataset, we use an LLM-based autorater to classify instances as sufficient or not. Here, an *autorater* is a model that evaluates instances based on a property, e.g., a sufficient context autorater.

Using our sufficient context autorater, we uncover new insights into LLM behavior and into existing benchmark datasets. First, we find models generate incorrect answers on a non-trivial fraction of instances that have sufficient context to answer the query. In other words, open-book QA cannot be solved by improving retrieval alone. Second, when given instances without sufficient context, models tend to hallucinate more than they abstain, especially for multi-hop questions. This finding complements prior work, which shows that LLMs are not robust to noisy retrieval [Yoran et al., 2024, Wu et al., 2024]. Third, models generate correct answers in many cases, even when the provided context is insufficient. Surprisingly, this remains true after we filter out questions that the model answers correctly in a closed-book (w/o RAG) setting. Together, our analysis deepens our understanding of RAG systems by revealing nuances in how models generate responses with retrieval.

As a final contribution, we explore ways to use sufficient context labels to reduce model hallucinations. We implement a new selective generation framework that improves accuracy. We

use a smaller intervention model to determine when the model generates or abstains, providing a controllable trade-off. Moreover, we can combine our method with any LLM, including proprietary models like Gemini and GPT. Our main result is that using sufficient context as an additional signal leads to much higher accuracy over the fraction of answered queries, for most coverage levels and across multiple models/datasets. We also find that fine-tuning open-source models with sufficient context information does not easily reduce the hallucination rate. Instead, for Mistral 3, fine-tuning can lead to a higher abstention rate at the cost of fewer correct answers. Key findings and the prompts used in our autorater analysis are available on our github.

To summarize, our main contributions are

1. We define the notion of sufficient context, unifying existing work on relevance for RAG systems. Then, we design a sufficient context autorater (achieving 93% accuracy), enabling us to label instances scalably and to analyze model responses with or without sufficient context.

2. Our analysis leads to several new findings about retrieval-augmented model performance. One takeaway is that SOTA LLMs output correct responses 35–62% of the time with insufficient context. Hence, intervention strategies to increase accuracy should not solely rely on sufficiency.

3. Building on our findings above, we develop an efficient and general method for selective generation, using both confidence and sufficient context signals. Our method improves the fraction of correct answers (among total model responses) by up to 2–10% for Gemini, GPT, and Gemma.

3.2 Related Work

Many papers have shown that reaping the benefits of RAG (e.g., better factuality) will require a deeper understanding of how LLMs respond to variations in the queries and provided context [Asai et al., 2024, Fan et al., 2024, Ram et al., 2023, Rau et al., 2024]. We review two

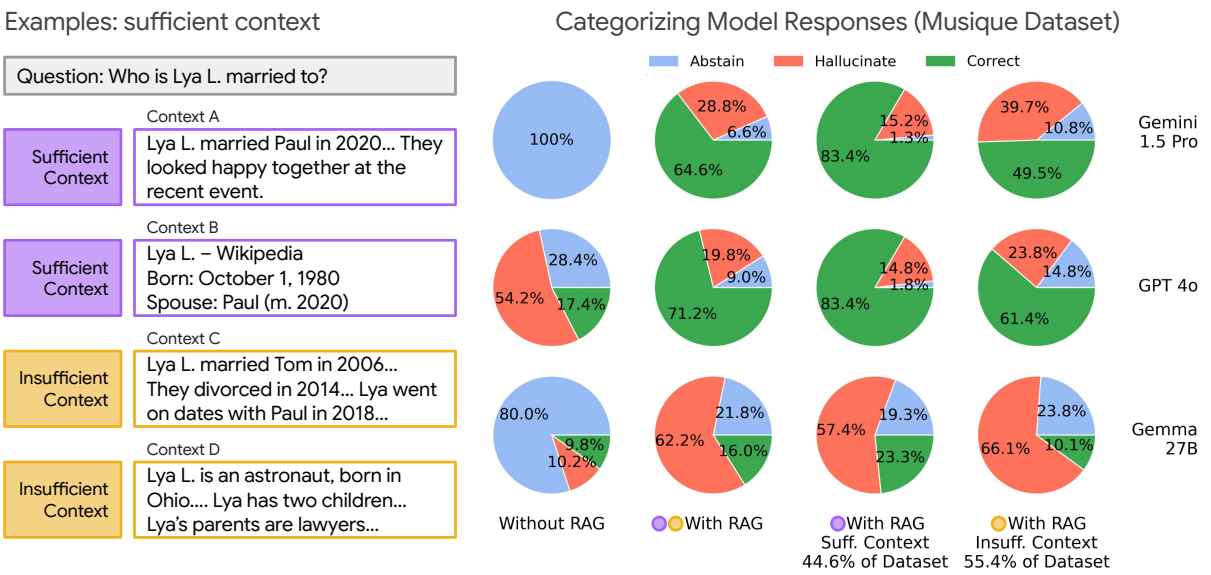


Figure 3.1. New insights into RAG systems by looking at whether instances have sufficient context. On the left, we show examples of sufficient context; on the right, a breakdown of model responses on the Musique dataset. Adding RAG improves the percentage of correct answers. Unfortunately, with RAG, models hallucinate more than abstain, and the insufficiency of the context does not account for this major issue. Also, standard datasets have many instances with insufficient context (here, 55.4%). We include results for other datasets (FreshQA, HotPotQA) in Appendix C.2, showing similar trends.

main areas. First, much work has evaluated RAG systems with poor retrieval, uncovering cases where LLMs are led astray by irrelevant context. Another line of study aims to reduce LLM hallucinations in RAG settings.

(Ir)relevant Context. Prior studies uncover a lack of robustness to imperfect retrieval. However, these studies vary in terms of how they evaluate retrieval quality, without anchoring to a precise "relevance" definition. Shi et al. [2023a] adds sentences to math questions (based on GSM8K), which should not impact the answer at all (and GSM8K is designed to have sufficient context by definition). Xie et al. [2024] looks at having counter-memory context, by either replacing the entity name with an erroneous one or using an LLM to generate a synthetic context supporting the erroneous entity. Ret-Robust [Yoran et al., 2024] trains a model to be robust to irrelevant context, with an NLI-based entailment to determine relevance, and only uses the relevance scores to influence the training mixture of relevant vs. irrelevant documents. Wu

et al. [2024] looks at questions where the LLM gets the answer correct without retrieval and is non-robust to changes in the retrieval. Multiple methods use a model to predict relevance scores (as part of a larger pipeline), without calibration to a formal definition [Wang et al., 2024a, Zhou et al., 2024], including for iterative retrieval [Jiang et al., 2024, Yan et al., 2024]. In terms of analysis studies, Cuconasu et al. [2024] distinguishes golden and relevant documents, but simply uses "does not contain the answer" as a proxy for irrelevant context.

Reducing Hallucinations. There have also been efforts to improve RAG factuality on open-book QA tasks [Asai et al., 2023, Mineiro, 2024, Simhi et al., 2024, Wang et al., 2024b, Zhang et al., 2024b]. The main theme is to improve both the generation and retrieval quality, often by fine-tuning one or more components. Also, since RAG leads to very long contexts, another issue that arises is the "lost in the middle" problem [Hsieh et al., 2024, Liu et al., 2024, Yu et al., 2024]. These works start with the premise that the provided query/context should be precisely answerable by the LLM, and hence, only analyze their findings in the sufficient context scenario. Independent of RAG, many papers have studied interventions and tools for calibrating LLM confidence in their responses [Chuang et al., 2024, Kadavath et al., 2022, Yin et al., 2023, Zhang et al., 2024a] and performance across disaggregated subsets of data [Paes et al., 2022, Joren et al., 2023].

3.3 Sufficient Context

At a high level, we aim to classify input instances based on whether the context contains enough information to answer the query. We split possible contexts into two cases: (1) **Sufficient Context.** The context is sufficient to answer the query if it contains all the necessary information to provide a definitive answer. (2) **Insufficient Context.** Otherwise, a context is insufficient. A context may also be insufficient if the query requires specialized knowledge that is not provided in the context or if the information in the context is incomplete, inconclusive, or contradictory. In this section, we more thoroughly discuss sufficient context. Then, we show how to accurately

and scalably label instances.

3.3.1 Definition of Sufficient Context

We first set some notation for a generic open-domain question-answering setting following Trivedi et al. [2020]. Consider a dataset D with instances of the form $q = (Q, C; A)$, where Q is the query and C is the context that consists of a set of facts. At inference time, we also consider instances $q' = (Q, C)$ without the ground truth answer, where the goal is to predict an answer A' from Q, C, and the model's parametric knowledge. To measure correctness, we compare A' and A, where there are many options such as exact match, F1 score, or an LLM-based assessment of answer sameness (we use an LLM). Using this notation, we can now define our notion of sufficient context.

Definition (Sufficient Context). An instance $q' = (Q, C)$ has sufficient context if and only if there exists an answer A' such that A' is a plausible answer to the question Q given the information in C.

To understand this, we can build on the Attributable to Identified Sources (AIS) framework [Rashkin et al., 2023]. Entailment via AIS answers a slightly different question. Namely, given an instance $q = (Q, C; A)$, the entailment objective is to determine the truth value of the proposition: The answer to the question Q is A given the information in C. The key difference between entailment and sufficient context is that for sufficient context, we do not presuppose that we have the answer A' in advance, only that such an answer exists. Finally, we only consider "plausible" answers, where we mean that A' could be an answer to the question Q. For example, if the question asks about a person's birthplace, then A' should be a location. We note that this allows for the possibility that the context contains an *incorrect* answer to the question. This is a key requirement, because (i) we would like to be able to use signal from sufficient context at inference time, where we do not have ground truth answers (see Section 3.5.1), and (ii) we hope to elucidate findings that are robust to ground truth label noise.

Remark 1 (Multi-hop queries). In most benchmark datasets (e.g., Musique, HotPotQA),

models are expected to be able to do multi-hop reasoning up to four hops, in which they must combine facts to form the answer. However, they should not infer connections that are not in the context. For example, if "Bob's mother was born in New York" then this does not suffice to say Bob was born in New York. But, if the context also says "Bob's mother is Alice..." and "... all of Alice's children were born in New York" then this instance has sufficient context.

Remark 2 (Ambiguous queries). If the query is ambiguous, then the context is sufficient if and only if (i) the context can disambiguate the query and (ii) the context provides an answer to the disambiguated query. For example, the question could be "What sport does Mia play?" and the context could contain both "Mia, from New York, plays basketball..." and "... Mia, from California, plays volleyball." This is sufficient because if the query is referring to either Mia from New York or Mia from California, then the context can answer the question.

Remark 3 (Ambiguous contexts). Assume the context contains multiple plausible answers to the query. Then it is sufficient if and only if it also provides enough information to distinguish between queries that would lead to each answer. For example, if the question is "What country does Ali live in?" and the context is "Ali lives in Paris" then this instance does not have sufficient context because it is not clear if Ali lives in Paris, France, or Paris, Texas, USA. If the context further contains "This weekend, Ali took the train from Paris to Marseille." Then this becomes sufficient because it is almost certain that Ali lives in France, as one cannot take a train from Texas to France.

3.3.2 Sufficient Context AutoRater

Next, we consider automating the task of labeling whether instances have sufficient context or not. We investigate two questions: (1) Can today's models achieve high accuracy on a challenging, human-annotated dataset? (2) How does an entailment model compare to general-purpose LLMs? To answer these questions, we evaluate methods on human-labeled data. Table 3.1 shows that Gemini 1.5 Pro can serve as an accurate autorater to label instances in terms of sufficient context. It achieves 93% accuracy, outperforms other methods, and operates without

Table 3.1. Sufficient Context AutoRater. Evaluating model ability to classify sufficient context on a gold-labeled dataset of 115 `(query, context; answer)` instances. Gemini 1.5 Pro (1-shot) performs the best, while FLAMe can be a cheaper alternative. TRUE-NLI and Contains GT need ground truth (GT) answers, while others only use `(query, context)`. Best in column in bold.

Metrics: / **Methods**	F1 Score	Accuracy	Precision	Recall	No GT
Gemini 1.5 Pro (1-shot)	**0.935**	**0.930**	0.935	**0.935**	✓
Gemini 1.5 Pro (0-shot)	0.878	0.870	0.885	0.871	✓
FLAMe (fine-tune PaLM 24B)	0.892	0.878	0.853	**0.935**	✓
TRUE-NLI (fine-tune T5 11B)	0.818	0.826	**0.938**	0.726	
Contains GT	0.810	0.809	0.870	0.758	

needing a ground truth answer.

Sufficient Context Labeled Dataset. Using the above definition, we construct gold labels for each `(query, context)` pair. We did not use ground truth answers or model responses. For the instances, we sample a total of 115 instances (queries, contexts, and answers) from standard benchmarks (PopQA, FreshQA, Natural Questions, EntityQuestions). We design the dataset to be very challenging, including single- and multi-hop questions, as well as adding highly related information in the context, even if it is not sufficient (e.g., a named entity from the question often appears in the context). We evaluate methods' abilities to classify sufficient context (binary labels).

Methods: Operating on Query-Context pairs. We use Gemini 1.5 Pro with either instructions (0-shot) or both instructions and a 1-shot example, held out from our dataset. FLAMe 24B is a general autorater model [Vu et al., 2024], but it has a small context window. For FLAMe, we divide the contexts into 1600 token chunks and ask whether each chunk is sufficient. If any chunk is labeled sufficient, we consider the instance to have sufficient context; otherwise, it's labeled as insufficient. We design prompts (in Appendix C.3) for both models based on the sufficient context definition above.

Methods: When a Ground Truth (GT) Answer is Available. For two baselines (TRUE-NLI, Contains GT), we use answers as an additional input to classify sufficient context.

TRUE-NLI is a fine-tuned entailment model [Honovich et al., 2022] that checks if the context entails one of the GT answers. Contains GT checks if a GT answer appears in the context. Comparing to entailment is particularly interesting because if a given answer A is entailed by (Q,C), then the context is also sufficient. On the other hand, the reverse is not true, since the answer A is only one possible choice for A'. As one consequence, if the ground truth answer A is incorrect, then it may not be entailed by the context. This happens when a named entity is ambiguous (e.g., two people with the same name), and the GT answer is based on one of the people while the context describes the other.

Results. Table 3.1 shows that Gemini 1.5 Pro (1-shot) performs the best overall in terms of F1 score and accuracy. As expected, TRUE-NLI has higher precision and lower recall: it measures entailment, which implies sufficient context. FLAMe outperforms TRUE-NLI in F1 and accuracy, but lags behind Gemini (1-shot), likely because it is a smaller model. The Contains GT method works surprisingly well, indicating that the presence of a ground truth answer correlates with context sufficiency.

Discussion. In real-world scenarios, we cannot expect candidate answers when evaluating model performance. Hence, it is desirable to use a method that works using only the query and context. Among these methods, Gemini 1.5 Pro (1-shot) has high accuracy and balanced precision and recall. Therefore, we use it in Section 3.4 as our main method for analyzing datasets and model responses. Later, in Section 3.5.1, we use FLAMe as a computationally efficient autorater to provide a signal for selective generation. Our comparison with TRUE-NLI and Contains GT confirms that classifying sufficient context is a different (and more complex) task than determining entailment.

3.4 A New Lens on RAG Performance

We set out to understand RAG performance by looking at sufficient context. We first analyze datasets (Section 3.4.1), then we investigate model performance with/without sufficient

context (Section 3.4.2). We qualitatively discuss cases where insufficient context leads to correct model responses (Section 3.4.3).

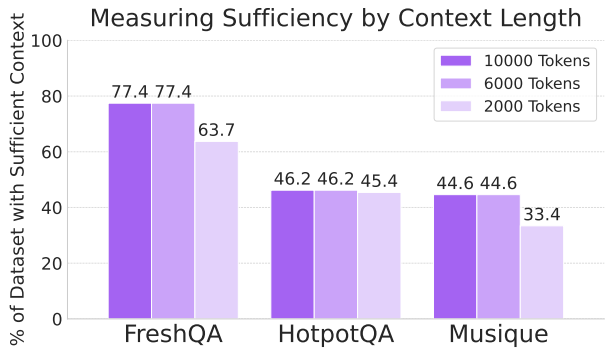


Figure 3.2. We compare the % of instances that our autorater labels as sufficient across datasets, either with the first 10k, 6k, or 2k tokens of the provided sources. FreshQA has hand-curated URLs that support the answers and exhibit high sufficient context. HotPotQA and Musique have lower sufficient context (and even lower with 2000 tokens). We use 6000 token contexts in the remainder.

3.4.1 Do Benchmark Datasets Have High Sufficient Context?

We introduce the datasets that we use for our analysis. Then, we investigate the percentage of instances in these datasets that have sufficient context (according to our autorater). For our study, we do not aim to optimize the retrieval methods (which could increase the sufficient context percentage). This is not the focus of our work, as we wish to understand how models perform with or without sufficient context. Having a mix of both is inevitable in generic RAG systems.

Datasets. We consider **FreshQA**, **Musique-Ans**, and **HotpotQA** as a representative spread of open book QA datasets. FreshQA [Vu et al., 2023] evaluates time-sensitive information and has up-to-date URLs that should support an answer to the queries, which we use to construct the `context` (see Appendix C.1.3 for details on the retrieval). We use the 'True Premise' setting (452 instances), skipping 'False Premise' questions that mislead by design. Musique-Ans [Trivedi et al., 2022] is a multi-hop QA benchmark, created by composing two to four single-hop interconnected questions. Here, 'Ans' is the standard 'answerable' subset. Musique instances

have 20 supporting text snippets as sources, which we use as the context. HotPotQA [Yang et al., 2018] is a Wikipedia-based QA dataset, with single- and multi-hop questions. The corpus is a large set of snippets; we retrieve the top 5 as the context (via REPLUG [Shi et al., 2023b] from FlashRAG [Jin et al., 2024]). We randomly sample 500 instances from the development sets of Musique-Ans and HotPotQA for evaluation.

Sufficient Context % of Datasets. Figure 3.2 shows the fraction of instances that our autorater classifies as having sufficient context. We explore three context lengths, ranging from a maximum of 2000 to maximum of 10000 tokens. The motivation behind this is to assess if there is a large change in sufficient context if we were to simply truncate the retrieval (e.g., for models that have small context windows). In general, we see a modest difference from 2000 to 6000 token limit, but effectively none from 6000 to 10000 tokens. FreshQA has the highest sufficient context percentage, which makes sense as the context comes from oracle supporting documents. The lower sufficient context in Musique is perhaps surprising, given that the retrieval is fixed as part of the dataset. From the results in Figure 3.2, we truncate at 6000 tokens for all three datasets in the remainder of the paper.

3.4.2 Initial Findings Based on Sufficient Context

In general, the ideal behavior for a language generation model is to answer questions correctly when possible and to otherwise abstain. RAG seeks to move models towards this desired behavior, such that the provided context shifts hallucinations to correct answers, or to abstentions if needed. We analyze several cases to assess how far we are from this ideal trade-off.

Experimental Set-up and LLMEval. We employed a basic chain of thought (CoT) prompting approach, with the prompt structure and further information detailed in Appendix C.3.4. We then processed the outputted answers to identify matches between the response and any of the ground truth answers. Responses where a clear, correct match could not be determined were processed through the LLMEval pipeline using a zero-shot approach, with the prompt based on Krishna et al. [2024] (see Appendix C.3.3). Then, for each example, we can rate it as "correct"

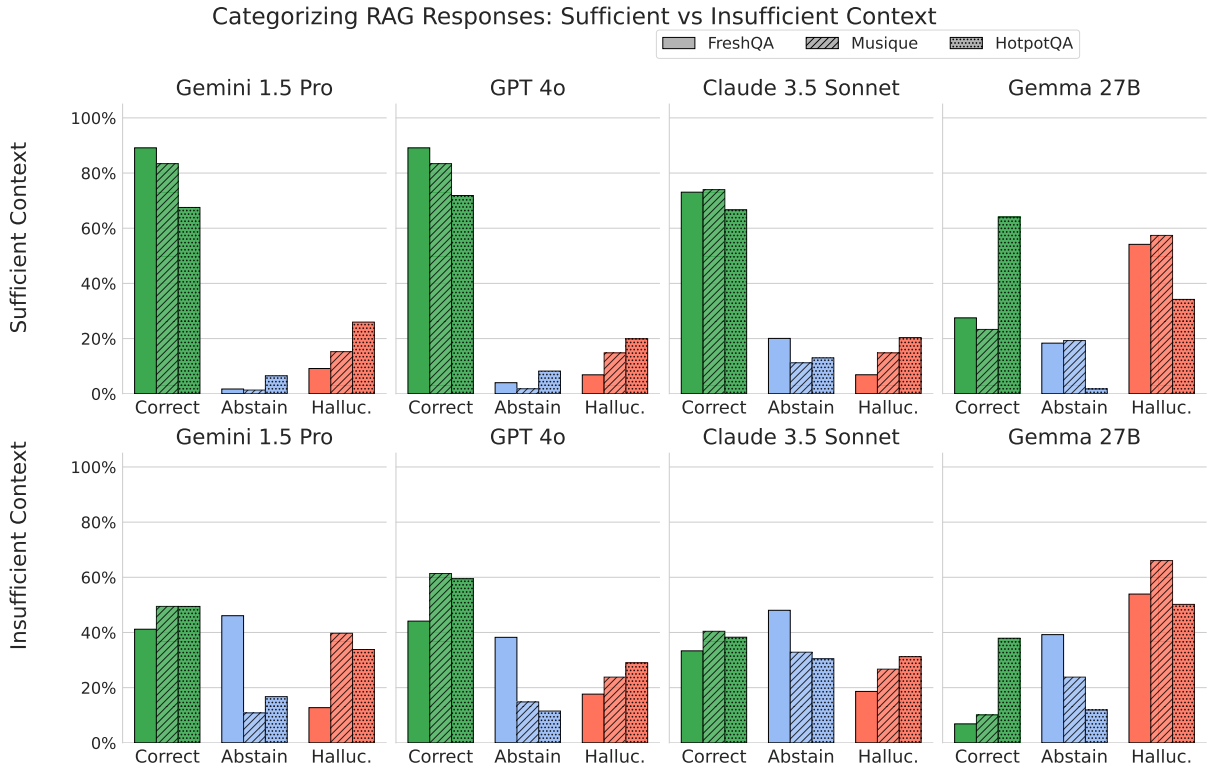


Figure 3.3. Model Performance on Datasets Stratified by Sufficient Context. Given sufficient context, models have a higher correct percentage on these challenging datasets. Performance drops, but the models are still able to answer a large portion of questions correctly without sufficient context. One prevailing issue is that all models hallucinate rather than abstain in many cases, with insufficient context. The smallest model, Gemma 27,B struggles to avoid hallucinations given insufficient context.

or "abstain" or "hallucinate" depending on the LLMEval output. We use an LLM for evaluation instead of checking for an exact match because it is more robust to syntactic variations. See Appendix C.2.3 for details and examples.

Models Abstain Less with RAG. While overall performance improves with RAG, the introduction of additional context paradoxically reduces the model's ability to abstain from answering when appropriate. Without RAG, Claude 3.5 Sonnet abstains on 84.1% of questions, while With RAG, the fraction of abstentions drops to 52%. Similarly, GPT 4o's abstention fraction moves from 34.4% to 31.2% and Gemini 1.5 Pro's drops from 100% to 18.6%. This phenomenon may arise from the model's increased confidence in the presence of any contextual information, leading to a higher propensity for hallucination rather than abstention.

Models Hallucinate with Both Sufficient and Insufficient Context. Considering Figure 3.3, models generally achieve higher accuracy with sufficient context (higher **green bars**, top row) than without sufficient context (lower **green bars**, bottom row). However, looking at each row separately, we discover several findings. First, in the sufficient context case (top row), we see that models hallucinate more than they abstain (**red bars** are higher than **blue bars**, usually). The trend holds across all three datasets. Moving to insufficient context (bottom row), we find a different distribution of model responses, with more abstentions and hallucinations. This tendency varies notably across different models. For instance, Claude abstains more (higher **blue bars**) with insufficient context, but answers fewer questions correctly (lower **green bars**) than Gemini and GPT. These differences underscore the potential for improvement in both retrieval and reasoning capabilities. Overall, Gemma has many more hallucinations (higher **red bars**) than the other models, except for HotPotQA, where we attribute the higher accuracy to the smaller retrieved contexts.

3.4.3 Qualitatively analyzing responses with insufficient context

One curious observation in our analysis is the ability of models to sometimes provide correct answers even when presented with insufficient context. For example, from Figure 3.3, all three models are able to correctly answer upwards of 35% of instances with insufficient context on HotpotQA. A natural assumption is that the models already know the answer from pre-training, and they can generate a correct response from parametric memory. However, this only explains part of the story.

Looking deeper, we provide a qualitative categorization in Table 3.2 of instance types where our autorater labels an instance as insufficient context, while the LLM evaluator marks the model answer as correct. For example, one type accounts for when the provided context is not sufficient to answer the query, but it bridges gaps in the model's knowledge. Another type is when the retrieved information clarifies ambiguities inherent in the question (without answering the question). Finally, we also have the times where either the autorater or the evaluator model

Table 3.2. Qualitative Analysis of Correct Answer & Insufficient Context. Examining model responses across datasets, we identify common cases where the model generates a correct answer even though our autorater labels the instance as insufficient. We categorize such instances into eight types, as well as provide examples. Given that models are also correct on many questions in the closed-book setting, we believe this mostly explains the 35–62% correct rate with insufficient context.

Instance type	**Why model may be correct**	**Example**
Yes/No question	50% chance of correct	**Q:** Is there a total eclipse in the United States this year?
Limited choice	Some chance of correct	**Q:** Which band has more members, Chvrches or Goodbye Mr. Mackenzie?
Multi-hop: fragment	Use parametric inference	**Q:** Who did the original voice for the character whose series Mickey's Safari in Letterland is from? *Context says Mickey's Safari is a video game and Walt Disney voices Mickey Mouse in cartoons. Must infer the game is in the Mickey Mouse series.*
Multi-hop: partial	Use parametric knowledge	**Q:** Claudine's Return starred the actress who played which role on "Married...with Children"? *Context lists actresses but not their roles in "Married...with Children". Must know extra facts.*
Too many hops	Execute complex reasoning	**Q:** How many cyclists have won all three of women's cycling Grand Tours equivalents in the same year? *Context requires cross-referencing lists of events and lists of winners while tracking winners by year.*
Ambiguous query	Guess right interpretation	**Q:** Who is the spouse of a cast member from King of the Mountain? *Context has many cast members and query/context do not specify which spouse to answer about.*
Rater error	Mislabel insuff. or correct	—
Closed-book correct	Known from pre-training	—

makes an error. We note that our analysis expands on prior work by Yoran et al. [2024], who also find a large fraction of cases where the model is correct with RAG (but not without) even though the context does not contain the answer.

We additionally investigated cases where the autorater labels an instance as having sufficient context while the LLM evaluator marks the answer as incorrect. One source of these discrepancies occurs when the ground truth answer conflicts with the answer provided in the source. This represents a key difference from methods that measure entailment, where context is evaluated relative to a specific ground truth answer (see Table 3.1). Another source of errors arises when the autorater correctly identifies that the necessary information is present, but the model fails to properly compose the information (e.g., in multihop questions or questions

requiring arithmetic). In a substantial number of cases, however, determining the source of the error proves challenging.

3.5 Techniques to Reduce Hallucinations with RAG

From our previous analysis, we have seen that models may hallucinate rather than abstain and that this happens more with RAG than in a closed-book setting. A natural next question is whether we can prompt or fine-tune a model to perform closer to the ideal case. Can we steer the model to either output the correct answer or abstain, while hallucinating an incorrect answer as little as possible?

3.5.1 Selective RAG Using Sufficient Context Signal

One simple solution to improving RAG performance would be to use the sufficient context autorater to abstain given insufficient context. However, this heavy-handed approach can lower overall performance, since all models answer some questions correctly even with insufficient context, as described in Table 3.2 and demonstrated in Figure 3.3. Instead, we propose a method for combining the sufficient context autorater outputs with model self-rated confidence scores to tune a selective accuracy-coverage trade-off, where "coverage" denotes the portion of inputs on which the model does not abstain. Specifically, we use these signals to train a simple linear model to predict hallucinations, and then use it to set coverage-accuracy trade-off thresholds.

This mechanism differs from other strategies for improving abstention in two key ways. First, because it operates independently from generation, it mitigates unintended downstream effects, whereas strategies like fine-tuning to improve abstention can inadvertently worsen performance on certain inputs (see Section 3.5.2). Second, it offers a *controllable* mechanism for tuning abstention, which allows for different operating settings in differing applications, such as strict accuracy compliance in medical domains or maximal coverage on creative generation tasks.

Abstention Signals

We utilize two main signals for abstention: the self-rated probabilities as in Li et al. [2024], Kadavath et al. [2022] and the sufficient context autorater. For the self-rated probabilities, we use two strategies: P(True) and P(Correct). P(True) requires sampling answers from the model multiple times, and then prompting the model multiple times to label each model as correct or incorrect, resulting in a final probability of correctness associated with each question, as in Kadavath et al. [2022]. For proprietary models, where extensive querying is prohibitively expensive, we use P(Correct) instead. We adapt the probability-generating prompt from Li et al. [2024] to obtain the model's response and its estimated probability of correctness. For the sufficient context signal, we use the binary label from an autorater. We hypothesize that combining these signals should lead to more effective abstention, particularly in cases where the context is insufficient.

Methods. We calculate P(True) by sampling 20 responses for each question and querying the model 5 times to evaluate whether the answer is correct or incorrect (without using the ground truth) as in Kadavath et al. [2022]. For P(Correct), the prompt requests the most likely and second most likely answers along with their probabilities. We use string matching to extract the response and self-predicted probability, keeping the one with the highest probability. To determine sufficient context, we use FLAMe, a small and efficient model for determining the sufficient context label. We divide the retrievals into chunks of 1600 tokens to fit in the context window and label the context as sufficient if any of these chunks are sufficient.

We combine the binary sufficient context label with the self-rated answer probability (P(True) for open-source models or P(Correct) for proprietary models) in a simple logistic regression model to predict hallucinations with 100 iterations of random hyperparameter search. At inference time, we use the logistic regression model scores to threshold the outputs, abstaining when the score is below a chosen threshold as in Joren et al. [2024]. We measure the added value for selective accuracy of the sufficient context signal (**purple line** in Figure 3.4) by comparing it with the model self-rated confidence alone (**gray line**).

Results. We find that our approach leads to a better selective accuracy-coverage trade-off compared to using model confidence alone. In particular, see gains of over 10% for Gemma 27B on HotpotQA in the highest accuracy regions, and gains of over 5% for Gemini 1.5 Pro on the same dataset near the 70% coverage region. These gains are less pronounced on datasets with lower overall accuracy, such as when using Gemma 27B on Musique. In this scenario, the low overall performance (18.4%) likely means that most of the predictive gains are seen by using the self-rated confidence to predict errors for the majority of samples. As a result, there is no added benefit from the sufficient context signal.

Discussion. As expected, we see a downward trend in which higher coverage leads to lower selective accuracy for both methods. We conclude that the selective generation mechanism with sufficient context has an added benefit for accuracy-coverage trade-offs compared to self-rated confidence alone. As a prerequisite for our method, models should have a non-trivial accuracy on sufficient and insufficient context instances. Then, intuitively, we can prioritize answering questions the model is likely to get right before those the model struggles with. While ordering examples is impossible in real settings, we can estimate a coverage level and use a threshold to choose when to answer.

3.5.2 Fine-tuning

We also consider fine-tuning models to increase their ability to abstain instead of outputting an incorrect answer. To do so, we train the models with some examples that contain "I don't know" instead of their original ground truth answer. The intuition here is that training explicitly on such inputs could encourage the model to abstain instead of hallucinating. We also consider multiple settings, such as only changing the answers when the example has insufficient context. We present full details in Appendix C.2.1. The main takeaways are that fine-tuned models (i) have a higher rate of correct answers in many cases, but (ii) still hallucinate quite often and more than they abstain. Overall, it is likely possible to use fine-tuning to steer the model towards better abstention and correctness, but more work is needed to develop a reliable

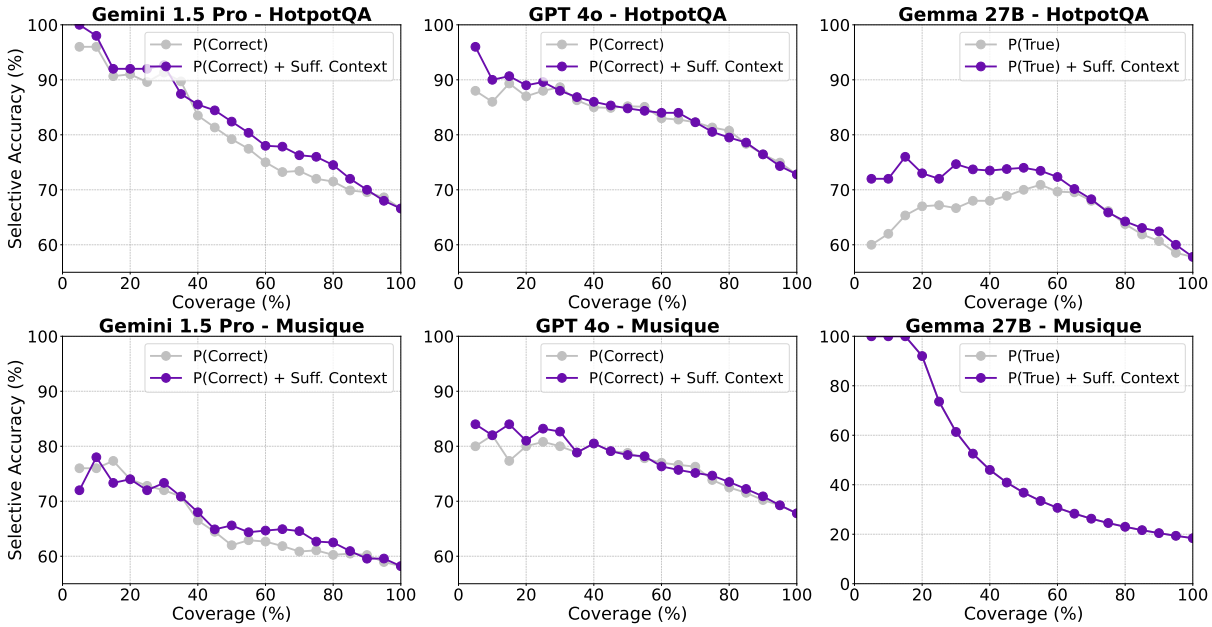


Figure 3.4. Selective Generation: Coverage vs. Selective Accuracy. For selective generation, we use a linear combination of sufficient context and self-rated confidence (**purple**) or confidence alone (**gray**). The x-axis shows coverage (% of questions answered); the y-axis shows accuracy at each coverage (# correct / # answered). The combined approach matches or outperforms the baseline confidence-only method, especially on HotpotQA, where our method improves accuracy for most coverages. For Gemma 27B on Musique, the methods are identical (coeff. for suff. context is 0).

strategy that can balance these objectives.

3.6 Conclusion

Our work provided a new lens on LLM responses in RAG systems centered around our notion of sufficient context. We constructed a sufficient context autorater, which enabled scalable insights into model performance on different types of instances. Our analysis revealed that even with sufficient context, LLMs frequently hallucinate answers. We also found, surprisingly, many cases where a model will output a correct answer with access to only insufficient context. Qualitatively, we categorized such instances, leading to a fuller picture of ways context can be useful. Finally, we demonstrated a general-purpose selective generation method, which applies to Gemini, GPT, and Gemma, and can reduce hallucinations by 2–10% on queries that the model

answers.

Limitations. Our analysis focuses on QA datasets, but summarization tasks also utilize context, which may or may not be sufficient. For example, models may behave differently on the prompt "Summarize the reviews of 5-star hotels in Mallorca" depending on whether the context mentions the hotel reviews, whether they are for 5-star hotels, etc. Another shortcoming is an exploration of how often different retrieval methods lead to sufficient context. Also to achieve the best performance, We could have used our autorater to iteratively judge whether to retrieve more or answer the question.

Future Work. One direction is a fine-grained sufficient context autorater, which outputs a score instead of a binary label. This could be useful for ranking contexts after the retrieval step. Another direction is to extend the definition of sufficient context to multi-modal RAG settings, such as for visual QA (images) or document QA (PDF files). Finally, our selective generation results suggest that there is room for improvement in reducing hallucinations by using auxiliary signals from the inputs.

Acknowledgments

Chapter 3, in full, is a reprint of the material as it appears in The Thirteenth International Conference on Learning Representations (ICLR 2025) Hailey Joren; Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. The dissertation author was the primary investigator and author of this paper. We thank Hrishikesh Garud, Vikram Gopali, Xun Sun, and Bruce Wang for annotating data. We thank Ranjay Krishna and Jacob Eisenstein for helpful discussions. We also thank Alyshia Olsen for help with the figure design and color palette. We thank the anonymous reviewers for their suggestions to improve the presentation.

Appendix A

Participatory Personalization

A.1 Supporting Material for Section 1.3

A.1.1 Enumeration Routine for Algorithm 1

We summarize the Enumeration routine in Algorithm 3. Algorithm 3 takes as input a set of group attributes $\mathscr{G}$ and a dataset $\mathscr{D}$ and outputs a collection of reporting interfaces $\mathbb{T}$ that obey ordering and plausibility constraints.

Algorithm 3. Enumerate All Possible Reporting Trees for Reporting Options $\mathscr{G}$

1: **procedure** VIABLETREES($\mathscr{G}, \mathscr{D}$)
2: **if** $\dim(\mathscr{G}) = 1$ **return** $[T_{\mathscr{G}}]$ *base case: we are left with only a single attribute on which to branch*
3: $\mathbb{T} \leftarrow [\,]$
4: **for** each group attribute $\mathscr{A} \in [\mathscr{G}_1, \ldots, \mathscr{G}_k]$ **do**
5: $T_{\mathscr{A}} \leftarrow$ reporting tree of depth 1 with $|\mathscr{A}|$ leaves
6: $\mathscr{S} \leftarrow \mathsf{ViableTrees}(\mathscr{G} \setminus \mathscr{A}, \mathscr{D})$ *all subtrees using all attributes except $\mathscr{A}$*
7: **for** Π in $\mathsf{ValidAssignments}(\mathscr{S}, \mathscr{A}, \mathscr{D})$ **do**: *each assignment is a permutation of $|\mathscr{A}|$ to leaves of $T_{\mathscr{A}}$*
8: $\mathbb{T} \leftarrow \mathbb{T} \cup T_{\mathscr{A}}.\mathsf{assign}(\Pi)$ *extends the tree by assigning subtrees to each leaf*
9: **end for**
10: **end for**
11: **return** $\mathbb{T}$, reporting interfaces for group attributes $\mathscr{G}$ that obey plausibility and ordering constraints
12: **end procedure**

The routine enumerates all possible reporting interfaces for a given set of group attributes $\mathscr{G}$ through a recursive branching process. Given a set of group attributes, the routine is called for each attribute that has yet to be considered in the tree Line 4, ensuring a complete enumeration. We note that the routine is only called for building Sequential systems since there is only one

possible reporting interface for Minimal and Flat systems.

Enumerating all possible trees ensures we can recover the best tree given the selection criteria and allows practitioners to choose between models based on other criteria. We generate trees that meet plausibility constraints based on the dataset, such as having at least one negative and one positive sample and at least s total samples at each leaf. In settings constrained by computational resources, we can impose additional stopping criteria and modify the ordering to enumerate more plausible trees first or exclusively (e.g., by changing the ordering of $\mathscr{G}$ or imposing constraints in VALIDASSIGNMENTS).

A.1.2 Assignment Routine for Algorithm 1

We summarize the routine for AssignModels procedure in Algorithm 4.

Algorithm 4. Assigning Models

procedure ASSIGNMODELS($T, \mathscr{M}, \mathscr{D}$)
  $Q \leftarrow [T.\text{root}]$ *initialize with the root of the tree, reporting group* $\varnothing$
  while Q is not empty **do**
    $\boldsymbol{r} \leftarrow Q.\text{pop}()$
    $\mathscr{M}_{\boldsymbol{r}} \leftarrow \text{ViableModels}(\mathscr{M}, \boldsymbol{r})$ *filter* $\mathscr{M}$ *to models that can be assigned to* $\boldsymbol{r}$
    $h^* \leftarrow \arg\min_{h \in \mathscr{M}_{\boldsymbol{r}}} \hat{R}_{\boldsymbol{r}}(h, \mathscr{D})$ *assign the model with the best training performance*
    $T.\text{set_model}(\boldsymbol{r}, h^*)$
    for $\boldsymbol{r}' \in T.\text{get_subgroups}(\boldsymbol{r})$ **do** *iterate through the children reporting groups of* $\boldsymbol{r}$
      $Q.\text{enqueue}(\boldsymbol{r}')$
    end for
  end while
  return T that maximizes gain for each reporting group
end procedure

Algorithm 4 takes as inputs a reporting tree T, a pool candidate models $\mathscr{M}$, and an assignment (training) dataset $\mathscr{D}$ and outputs a tree T that maximizes the gains of reporting group information. The pool of candidate models is filtered to viable models for each reporting group. Since the pool of candidate models includes the generic model h_0, each reporting group will have at least one viable model. We assign each reporting group the best-performing model on the training set and default to the generic model h_0 when a better-performing personalized model

is not found. We assign performance on the training set and then prune using performance on the validation set to avoid biased gain estimations.

A.1.3 Pruning Routine for Algorithm 1

We summarize the routine used for the PruneLeaves procedure in Algorithm 1. The PruneLeaves routine

Algorithm 5. Pruning Participatory Systems

procedure PRUNELEAVES($T, \mathscr{D}$)
  $Stack \leftarrow [T.\text{leaves}]$ *initialize stack with all leaves*
  repeat
    $\boldsymbol{r} \leftarrow Stack.pop()$
    $h \leftarrow T.\text{get_model}(\boldsymbol{r})$
    $h' \leftarrow T.\text{get_model}(\text{pa}(\boldsymbol{r}))$
    if not $\text{Test}(\boldsymbol{r}, h, h', \mathscr{D})$ **then** *test gains to see if parent model is as good as leaf model*
      $T.\text{prune}(\boldsymbol{r})$
    end if
    if $T.\text{get_children}(\text{pa}(\boldsymbol{r}))$ is empty **then** *consider pruning the parent if the parent has become a leaf*
      $Stack.\text{enqueue}(\text{pa}(\boldsymbol{r}))$
    end if
  until *Stack* is empty
  return T, reporting interface that ensures data collection leads to gain
end procedure

Algorithm 1 takes as input a reporting interface T and a validation sample $\mathscr{D}$, and performs a bottom-up pruning to output a reporting interface T that asks individuals to report attributes that are expected to lead to a gain. The pruning decision at each leaf is based on a hypothesis test that evaluates the gains of reporting for a reporting group on a validation dataset. This test has the form:

$$H_0 : R_{\boldsymbol{g}}(h) \leq R_{\boldsymbol{g}}(h') \quad \text{vs.} \quad H_A : R_{\boldsymbol{g}}(h) > R_{\boldsymbol{g}}(h')$$

This procedure evaluates the gains of reporting by comparing the performance of a model assigned at a leaf node h and a model assigned at a parent node h', which does not use the reported information. Here, the null hypothesis H_0 assumes that the parent model performs

as well as the leaf model – and thus, we reject the null hypothesis when there is sufficient evidence to suggest that reporting will improve performance in deployment. Our routine allows practitioners to specify the hypothesis test to compute the gains. By default, we use the McNemar test for accuracy Dietterich [1998] and the Delong test for AUC DeLong et al. [1988], Sun and Xu [2014]. In general, we can use a bootstrap hypothesis test [DiCiccio and Efron, 1996].

A.1.4 Greedy Induction of Sequential Reporting Interface

We present an additional routine to construct reporting interfaces for sequential systems in Algorithm 6. We include this routine as an alternative option that can be used to construct a reporting interface in settings where it may be impractical or undesirable to enumerate all possible reporting interfaces. The procedure results in a valid reporting interface that ensures gains. However, it does not guarantee an optimal tree in terms of maximizing the overall gain and does not allow practitioners to choose between reporting interfaces after training.

Algorithm 6. Greedy Induction Routine for Sequential Reporting Interfaces

procedure GREEDYTREE($\mathscr{R}$)
  $T \leftarrow$ empty reporting interface
  repeat
    for $\boldsymbol{r} \in \text{leaves}(T)$ **do**
      $\{\mathscr{A}_{\boldsymbol{r}}\} \leftarrow G_i : \boldsymbol{r}[i] = \varnothing$ *$\{\mathscr{A}_{\boldsymbol{r}}\}$ contains all heretofore unused attributes*
      $\mathscr{A}^* \leftarrow \arg\max_{\mathscr{A} \in \{\mathscr{A}_{\boldsymbol{r}}\}} \min_{\boldsymbol{r}' \in \boldsymbol{r}.\mathsf{split}(\mathscr{A})} \Delta_{\boldsymbol{r}'}(\boldsymbol{r}', \boldsymbol{r})$
      $\boldsymbol{r}.\mathsf{split}(\mathscr{A}^*)$ *Split on attribute that maximizes worst-case gain*
    end for
  until no splits are added
  return T, reporting interface that ensures gains for reporting each $\mathscr{R}$.
end procedure

Algorithm 6 takes as input a collection of reporting options $\mathscr{R}$ and outputs a single reporting interface using a greedy tree induction routine that chooses the attribute to report to maximize the minimum gain at each step. The procedure uses the reporting options to iteratively construct a reporting tree that branches on all of the attributes in $\mathscr{R}$. The procedure considers each unused attribute for each splitting point and splits on the attribute that provides the greatest minimum gain for the groups contained at that node.

A.2 Description of Datasets used in Section 1.4 – Experiments

We include additional information about the datasets used in Section 1.4.

Table A.1. Datasets used to fit clinical prediction models in Section 1.4. Here: n denotes the number of examples in each dataset; d denotes the number of features; $\mathscr{G}$ denotes the group attributes that are used for personalization; and $m = |\mathscr{G}|$ denotes the number of intersectional groups. Each dataset is de-identified and available to the public. The `cardio_eicu`, `cardio_mimic`, `lungcancer` datasets require access to public repositories listed under the references. The `saps` and `apnea` datasets must be requested from the authors. The `support` dataset can be downloaded directly from the URL below.

Dataset	Reference	Outcome Variable	n	d	m	$\mathscr{G}$
`apnea`	Ustun et al. [2016]	patient has obstructive sleep apnea	1,152	28	6	{`age`, `sex`}
`cardio_eicu`	Pollard et al. [2018]	patient with cardiogenic shock dies	1,341	49	8	{`age`, `sex`, `race`}
`cardio_mimic`	Johnson et al. [2016]	patient with cardiogenic shock dies	5,289	49	8	{`age`, `sex`, `race`}
`coloncancer`	Scosyrev et al. [2012]	patient dies within 5 years	29,211	72	6	{`age`, `sex`}
`lungcancer`	Scosyrev et al. [2012]	patient dies within 5 years	120,641	84	6	{`age`, `sex`}
`saps`	Allyn et al. [2016]	ICU mortality	7,797	36	4	{`age`, `HIV`}

`apnea`

We use the obstructive sleep apnea (OSA) dataset outlined in Ustun et al. [2016]. This dataset includes a cohort of 1,152 patients, where 23% have OSA. We use all available features (e.g., BMI, comorbidities, age, and sex) and binarize them, resulting in 26 binary features.

`cardio_eicu` & `cardio_mimic`

Cardiogenic shock is an acute condition in which the heart cannot provide sufficient blood to the vital organs [Hollenberg, 2003]. These datasets are designed to predict cardiogenic shock for patients in intensive care. Each dataset contains the same features, group attributes, and outcome variables for patients in different cohorts. The `cardio_eicu` dataset contains records for a cohort of patients in the Collaborative Research Database V2.0 [Pollard et al., 2018]. The `cardio_mimic` dataset contains records for a cohort of patients in the MIMIC-III [Johnson et al., 2016] database. Here, the outcome variable indicates whether a patient in the ICU with

cardiogenic shock will die while in the ICU. The features encode the results of vital signs and routine lab tests (e.g., systolic BP, heart rate, hemoglobin count) that were collected up to 24 hours before the onset of cardiogenic shock.

lungcancer

We consider a cohort of 120,641 patients who were diagnosed with lung cancer between 2004-2016 and monitored as part of the National Cancer Institute SEER study [Scosyrev et al., 2012]. Here, the outcome variable indicates if a patient dies within five years from any cause, and 16.9% of patients died within the first five years from diagnosis. The cohort includes patients from Greater California, Georgia, Kentucky, New Jersey, and Louisiana, and does not cover patients who were lost to follow-up (censored). Age and Sex were considered as group attributes. The features reflect the morphology and histology of the tumor (e.g., size, metastasis, stage, node count and location, number and location of nodes) as well as interventions that were administered at the time of diagnosis (e.g., surgery, chemo, radiology).

coloncancer

We consider a cohort of 120,641 patients who were diagnosed with colorectal cancer between 2004-2016 and monitored as part of the National Cancer Institute SEER study [Scosyrev et al., 2012]. Here, the outcome variable indicates if a patient dies within five years from any cause, and 42.1% of patients die within the first five years from diagnosis. The cohort includes patients from Greater California. Age and Sex were considered as group attributes. The features reflect the morphology and histology of the tumor (e.g., size, metastasis, stage, node count and location, number and location of notes) as well as interventions that were administered at the time of diagnosis (e.g., surgery, chemo, radiology).

saps

The Simplified Acute Physiology Score II (SAPS II) score predicts the risk of mortality of critically-ill patients in intensive care [Le Gall et al., 1993]. The data contains records of 7,797 patients from 137 medical centers in 12 countries. Here, the outcome variable indicates whether

a patient dies in the ICU, with 12.8% of patients dying. The features reflect comorbidities, vital signs, and lab measurements.

A.3 Results for Different Model Classes and Prediction Tasks

In this Appendix, we present experimental results for additional model classes and prediction tasks. We produce these results using the setup in Section 1.4.1, and summarize them in the same way as Table 1.2. We refer to them in our discussion in Section 1.4.2.

A.3.1 Logistic Regression for Ranking (AUC)

Table A.2. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc**. We show the performance of models and systems built using **logistic regression**.

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
apnea $n = 1152, d = 26$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Ustun et al. [2016]	Overall Performance	0.774	0.774	0.776	0.776	0.776	**0.851**	**0.851**
	Overall Gain	-0.002	-0.002	0.000	-0.000	0.000	**0.074**	**0.074**
	Group Gains	-0.002 – 0.002	-0.002 – 0.003	-0.002 – 0.002	-0.002 – 0.003	0.000 – 0.002	0.004 – 0.115	0.004 – 0.115
	Max Disparity	0.004	0.005	0.004	0.005	0.002	0.111	0.111
	Rat. Violations	**2**	**2**	**2**	**2**	0	0	0
	Imputation Risk	-0.002	-0.002					
	Options Pruned	0/6	0/6	0/12	0/12	5/7	4/12	4/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	100.0%	83.3%
cardio_eicu $n = 1341, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Pollard et al. [2018]	Overall Performance	0.864	0.863	0.863	0.862	0.865	0.966	**0.966**
	Overall Gain	0.002	0.001	0.000	-0.001	0.002	0.103	**0.103**
	Group Gains	-0.005 – 0.003	-0.010 – 0.010	-0.005 – 0.003	-0.010 – 0.010	0.000 – 0.003	0.010 – 0.180	0.010 – 0.180
	Max Disparity	0.009	0.019	0.009	0.019	0.003	0.170	0.170
	Rat. Violations	**3**	**3**	**3**	**3**	0	0	0
	Imputation Risk	-0.005	-0.010					
	Options Pruned	0/8	0/8	0/27	0/27	6/9	13/27	11/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	95.8%
cardio_mimic $n = 5289, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Johnson et al. [2016]	Overall Performance	0.881	0.881	0.882	0.880	0.881	**0.914**	**0.914**
	Overall Gain	0.000	0.000	0.002	-0.000	0.000	**0.034**	**0.034**
	Group Gains	-0.001 – 0.001	-0.001 – 0.001	-0.001 – 0.001	-0.001 – 0.001	0.000 – 0.001	0.008 – 0.057	0.008 – 0.057
	Max Disparity	0.002	0.002	0.002	0.002	0.001	0.049	0.049
	Rat. Violations	**3**	**3**	**3**	**3**	0	0	0
	Imputation Risk	-0.001	-0.001					
	Options Pruned	0/8	0/8	0/27	0/27	6/9	9/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	25.0%	100.0%	91.7%
coloncancer $n = 29211, d = 72$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	0.685	0.685	0.683	0.683	0.685	**0.700**	0.700
	Overall Gain	0.001	0.002	-0.000	-0.000	0.001	**0.016**	0.016
	Group Gains	-0.001 – 0.002	-0.001 – 0.001	-0.001 – 0.002	-0.001 – 0.001	0.000 – 0.001	0.001 – 0.021	0.001 – 0.021
	Max Disparity	0.003	0.002	0.003	0.002	0.001	0.020	0.020
	Rat. Violations	**3**	**2**	**3**	**2**	0	0	0
	Imputation Risk	-0.001	-0.002					
	Options Pruned	0/6	0/6	0/12	0/12	5/7	2/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	16.7%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	0.855	0.855	0.852	0.854	0.855	**0.861**	0.861
	Overall Gain	0.001	0.001	-0.002	0.000	0.001	**0.006**	0.006
	Group Gains	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	-0.000 – 0.000	0.000 – 0.000	0.001 – 0.012	0.001 – 0.012
	Max Disparity	0.001	0.001	0.001	0.001	0.000	0.011	0.011
	Rat. Violations	**2**	**2**	**2**	**2**	**1**	0	0
	Imputation Risk	-0.000	-0.000					
	Options Pruned	0/6	0/6	0/12	0/12	4/7	2/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	33.3%	100.0%	91.7%
saps $n = 7797, d = 36$ $\mathscr{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathscr{G}\rvert = 4$ groups Allyn et al. [2016]	Overall Performance	0.875	0.877	0.875	0.857	0.875	**0.960**	0.960
	Overall Gain	0.010	0.011	0.010	-0.008	0.009	**0.095**	0.095
	Group Gains	-0.000 – 0.016	-0.002 – 0.019	-0.000 – 0.016	-0.002 – 0.019	0.000 – 0.016	0.035 – 0.141	0.035 – 0.141
	Max Disparity	0.017	0.021	0.017	0.021	0.016	0.106	0.106
	Rat. Violations	**1**	**1**	**1**	**1**	0	0	0
	Imputation Risk	-0.000	-0.002					
	Options Pruned	0/4	0/4	0/9	0/9	1/5	2/9	3/9
	Data Use	100.0%	100.0%	0.0%	0.0%	75.0%	100.0%	87.5%

A.3.2 Random Forests for Decision-Making (Error)

Table A.3. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test error**. We show the performance of models and systems built using **random forests**.

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
apnea $n = 1152, d = 26$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Ustun et al. [2016]	Overall Performance	26.3%	26.0%	25.9%	27.4%	26.3%	**12.2%**	**12.2%**
	Overall Gain	1.5%	1.8%	1.9%	0.4%	1.5%	**15.6%**	**15.6%**
	Group Gains	-0.8% – 4.2%	0.4% – 3.8%	-0.8% – 4.2%	0.4% – 3.8%	0.0% – 4.2%	5.3% – 22.2%	5.3% – 22.2%
	Max Disparity	5.0%	3.4%	5.0%	3.4%	4.2%	16.9%	16.9%
	Rat. Violations	1	0	1	0	0	0	0
	Imputation Risk	-1.2%	-1.2%					
	Options Pruned	0/6	0/6	0/12	0/12	2/7	1/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	66.7%	100.0%	91.7%
cardio_eicu $n = 1341, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Pollard et al. [2018]	Overall Performance	18.6%	17.8%	18.2%	18.6%	18.4%	**5.7%**	6.0%
	Overall Gain	-0.2%	0.6%	0.2%	-0.2%	0.0%	**12.7%**	12.4%
	Group Gains	-3.5% – 1.4%	-2.2% – 3.0%	-3.5% – 1.4%	-2.2% – 3.0%	0.0% – 0.0%	6.0% – 14.9%	6.0% – 14.9%
	Max Disparity	4.9%	5.3%	4.9%	5.3%	0.0%	8.9%	8.9%
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-3.5%	-2.2%					
	Options Pruned	0/8	0/8	0/27	0/27	8/9	11/27	8/27
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	91.7%
cardio_mimic $n = 5289, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Johnson et al. [2016]	Overall Performance	19.9%	20.1%	19.9%	20.2%	19.6%	11.5%	**11.4%**
	Overall Gain	-0.3%	-0.5%	-0.3%	-0.6%	0.0%	8.1%	**8.1%**
	Group Gains	-1.1% – 1.3%	-1.3% – 0.5%	-1.1% – 1.3%	-1.3% – 0.5%	0.0% – 0.0%	1.0% – 14.9%	1.0% – 14.9%
	Max Disparity	2.4%	1.7%	2.4%	1.7%	0.0%	13.8%	13.8%
	Rat. Violations	5	6	5	6	0	0	0
	Imputation Risk	-1.1%	-1.3%					
	Options Pruned	0/8	0/8	0/27	0/27	8/9	6/27	5/27
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	87.5%
coloncancer $n = 29211, d = 72$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	37.2%	37.0%	37.2%	37.0%	37.0%	**35.9%**	35.9%
	Overall Gain	-0.2%	0.0%	-0.2%	-0.0%	0.0%	**1.0%**	1.0%
	Group Gains	-0.7% – 0.1%	-0.3% – 0.2%	-0.7% – 0.1%	-0.3% – 0.2%	0.0% – 0.0%	0.1% – 3.2%	0.1% – 3.2%
	Max Disparity	0.7%	0.5%	0.7%	0.5%	0.0%	3.1%	3.1%
	Rat. Violations	4	1	4	1	0	0	0
	Imputation Risk	-0.7%	-0.3%					
	Options Pruned	0/6	0/6	0/12	0/12	6/7	3/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	20.0%	20.2%	20.0%	20.3%	20.0%	**19.3%**	19.3%
	Overall Gain	0.1%	-0.1%	0.1%	-0.2%	0.1%	**0.8%**	0.7%
	Group Gains	-0.3% – 0.2%	-0.5% – 0.0%	-0.3% – 0.2%	-0.5% – 0.0%	0.0% – 0.2%	0.0% – 2.3%	0.0% – 2.2%
	Max Disparity	0.6%	0.5%	0.6%	0.5%	0.2%	2.3%	2.1%
	Rat. Violations	1	4	1	4	0	0	0
	Imputation Risk	-0.3%	-0.5%					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	1/12	3/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	83.3%
saps $n = 7797, d = 36$ $\mathscr{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathscr{G}\rvert = 4$ groups Allyn et al. [2016]	Overall Performance	14.1%	15.0%	14.1%	15.7%	13.9%	**9.8%**	**9.8%**
	Overall Gain	0.9%	-0.0%	0.9%	-0.7%	1.1%	**5.2%**	**5.2%**
	Group Gains	-0.8% – 3.4%	-0.5% – 0.3%	-0.8% – 3.4%	-0.5% – 0.3%	0.0% – 3.4%	0.0% – 16.4%	0.0% – 16.4%
	Max Disparity	4.2%	0.8%	4.2%	0.8%	3.4%	16.4%	16.4%
	Rat. Violations	1	1	1	1	0	0	0
	Imputation Risk	-0.8%	-0.7%					
	Options Pruned	0/4	0/4	0/9	0/9	2/5	1/9	1/9
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	75.0%	87.5%

A.3.3 Random Forests for Ranking (AUC)

Table A.4. Overview of performance, data use, and consent for all personalized models and systems on all datasets as measured by **test auc**. We show the performance of models and systems built using **random forests**.

		Static		Imputed		Participatory		
Dataset	Metrics	1Hot	mHot	KNN-1Hot	KNN-mHot	Minimal	Flat	Seq
apnea $n = 1152, d = 26$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Ustun et al. [2016]	Overall Performance	0.825	0.824	0.822	0.806	0.823	**0.944**	0.942
	Overall Gain	0.008	0.006	0.004	-0.012	0.005	**0.126**	0.124
	Group Gains	-0.004 – 0.009	-0.005 – 0.012	-0.004 – 0.009	-0.005 – 0.012	0.000 – 0.009	0.058 – 0.157	0.058 – 0.157
	Max Disparity	0.012	0.017	0.012	0.017	0.009	0.098	0.098
	Rat. Violations	2	3	2	3	0	0	0
	Imputation Risk	-0.004	-0.005					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	2/12	4/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	75.0%
cardio_eicu $n = 1341, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Pollard et al. [2018]	Overall Performance	0.896	0.896	0.897	0.886	0.894	**0.987**	0.987
	Overall Gain	0.003	0.003	0.004	-0.007	0.001	**0.094**	0.094
	Group Gains	-0.008 – 0.011	-0.005 – 0.011	-0.008 – 0.011	-0.005 – 0.011	0.000 – 0.004	0.010 – 0.132	0.010 – 0.130
	Max Disparity	0.020	0.016	0.020	0.016	0.004	0.122	0.120
	Rat. Violations	3	4	3	4	0	0	0
	Imputation Risk	-0.008	-0.005					
	Options Pruned	0/8	0/8	0/27	0/27	7/9	10/27	10/27
	Data Use	100.0%	100.0%	0.0%	0.0%	12.5%	100.0%	87.5%
cardio_mimic $n = 5289, d = 49$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}, \texttt{race}\}$ $\lvert\mathscr{G}\rvert = 8$ groups Johnson et al. [2016]	Overall Performance	0.884	0.883	0.884	0.881	0.885	**0.955**	0.954
	Overall Gain	0.000	-0.001	0.001	-0.002	0.001	**0.071**	0.071
	Group Gains	-0.005 – 0.006	-0.006 – 0.013	-0.005 – 0.006	-0.006 – 0.013	0.000 – 0.006	0.016 – 0.108	0.016 – 0.107
	Max Disparity	0.011	0.019	0.011	0.019	0.006	0.092	0.090
	Rat. Violations	3	7	3	7	0	0	0
	Imputation Risk	-0.005	-0.006					
	Options Pruned	0/8	0/8	0/27	0/27	5/9	6/27	6/27
	Data Use	100.0%	100.0%	0.0%	0.0%	37.5%	100.0%	83.3%
coloncancer $n = 29211, d = 72$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	0.684	0.682	0.681	0.680	0.683	**0.696**	0.696
	Overall Gain	0.002	0.000	-0.001	-0.002	0.001	**0.014**	0.014
	Group Gains	-0.002 – 0.004	-0.004 – 0.002	-0.002 – 0.004	-0.004 – 0.002	0.000 – 0.004	0.004 – 0.035	0.004 – 0.031
	Max Disparity	0.006	0.007	0.006	0.007	0.004	0.030	0.026
	Rat. Violations	0	0	0	0	0	0	0
	Imputation Risk	-0.002	-0.004					
	Options Pruned	0/6	0/6	0/12	0/12	3/7	2/12	5/12
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	75.0%
lungcancer $n = 120641, d = 84$ $\mathscr{G} = \{\texttt{age}, \texttt{sex}\}$ $\lvert\mathscr{G}\rvert = 6$ groups Scosyrev et al. [2012]	Overall Performance	0.849	0.849	0.848	0.849	0.848	**0.856**	**0.856**
	Overall Gain	0.002	0.001	0.001	0.001	0.000	**0.008**	**0.008**
	Group Gains	-0.001 – 0.003	-0.001 – 0.002	-0.001 – 0.003	-0.001 – 0.002	0.000 – 0.003	0.002 – 0.020	0.002 – 0.020
	Max Disparity	0.004	0.003	0.004	0.003	0.003	0.018	0.018
	Rat. Violations	1	1	1	1	0	0	0
	Imputation Risk	-0.001	-0.001					
	Options Pruned	0/6	0/6	0/12	0/12	2/7	1/12	2/12
	Data Use	100.0%	100.0%	0.0%	0.0%	66.7%	100.0%	91.7%
saps $n = 7797, d = 36$ $\mathscr{G} = \{\texttt{HIV}, \texttt{age}\}$ $\lvert\mathscr{G}\rvert = 4$ groups Allyn et al. [2016]	Overall Performance	0.921	0.922	0.922	0.906	0.921	**0.966**	**0.966**
	Overall Gain	0.003	0.004	0.003	-0.012	0.002	**0.048**	**0.048**
	Group Gains	-0.002 – 0.010	-0.002 – 0.013	-0.002 – 0.010	-0.002 – 0.013	0.000 – 0.010	0.009 – 0.109	0.009 – 0.109
	Max Disparity	0.012	0.015	0.012	0.015	0.010	0.100	0.100
	Rat. Violations	2	2	2	2	0	0	0
	Imputation Risk	-0.002	-0.002					
	Options Pruned	0/4	0/4	0/9	0/9	2/5	2/9	2/9
	Data Use	100.0%	100.0%	0.0%	0.0%	50.0%	100.0%	87.5%

A.4 Supporting Material for Performance Profiles

In the performance profiles, we measure the benefit of disclosure in terms of their expected performance gain and simulate the cost of reporting for each individual by sampling

their reporting cost from a uniform distribution – i.e., for each individual i, we sample c_i as $c_i \sim \text{Uniform}(0, \gamma),$ where $\gamma \in [0, 0.2]$. For each value of γ, we sample reporting costs ten times and average over the per-group performance error for each sampled cost.

Appendix B

Conceptual Safeguards

B.1 Omitted Proofs

Proposition 1. *Suppose that $\bar{y}$ is a calibrated probability prediction for y. Then any selective classifier $\varphi_\tau(\bar{y})$ that abstains when $\bar{y}$ has confidence below $1-\tau$ achieves accuracy at least $1-\tau$.*

Proof. We show that on examples for which the selective classifier does not abstain, accuracy is at least $1-\tau$. First, we use the Law of Total Probability to separate the cases where $\bar{y}$ is confident that $y=0$ versus $y=1$.

$$\begin{aligned}
&\Pr(y=\varphi_\tau(\bar{y}) \mid \varphi_\tau(\bar{y}) \neq \perp) \\
&= \Pr(\bar{y} \geq 1-\tau \mid \varphi_\tau(\bar{y}) \neq \perp) \Pr(y=1 \mid \bar{y} \geq 1-\tau) \\
&\quad + \Pr(\bar{y} \leq \tau \mid \varphi_\tau(\bar{y}) \neq \perp) \Pr(y=0 \mid \bar{y} \leq \tau)
\end{aligned}$$

Next, we use the definition of calibration, which states that $\Pr(y=1 \mid \bar{y}=t)=t$ for all $t \in [0,1]$, to write

$$\begin{aligned}
&\geq \Pr(\bar{y} \geq 1-\tau \mid \varphi_\tau(\bar{y}) \neq \perp)(1-\tau) + \Pr(\bar{y} \leq \tau \mid \varphi_\tau(\bar{y}) \neq \perp)(1-\tau) \\
&= (1-\tau)
\end{aligned}$$

□

B.2 Supporting Material for Experiments

B.2.1 Datasets

`melanoma` & `skincancer`

These datasets are derived from the Derm7pt repository Kawahara et al. [2018], which is de-identified and publicly available without patient information. We preprocess the dataset by splitting the original seven categorical image annotations (**pigment_network, streaks, pigmentation, regression_structures, dots_and_globules, blue_whitish_veil, vascular_structures**) into seventeen binary concepts. We consider two tasks: predicting `melanoma` and predicting `skincancer` (melanoma or basal cell carcinoma). We split the validation indices in the original dataset into a validation set and a hold-out test set for evaluation. We then balance the resulting classes by downsampling the majority class. To train the concept models, we augment the original training images with random color enhancements and random flipping to obtain 10x total training images.

`warbler`, `flycatcher`, `cubspecies` & `cubtypes`

These datasets are derived from the CUB 2011 dataset Wah et al. [2011]. We follow the same preprocessing described in Koh et al. [2020]. `warbler` classifies birds of type warbler and `flycatcher` classifies birds of type flycatcher. `cubspecies` and `cubtypes` are multiclass datasets for predicting bird species and bird types, respectively. To train the concept models, we augment the original training images with random color enhancements and random flipping to obtain 10x total training images.

`noisyconcepts`

The `noisyconcepts` datasets are synthetic datasets that we primarily use to evaluate performance changes with respect to the quality of concept detectors. We sample the examples

in these datasets $(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)$ from the following distribution:

$$
\begin{aligned}
x_1, \ldots, x_5 &= \text{Bernouilli}(0.7) \\
\xi_1, \xi_2, \xi_3 &= \text{Bernouilli}(p_\xi) \\
c_1 &= \text{parity}(x_1, x_2, x_4) \oplus \xi_1 \\
c_2 &= \text{parity}(x_1, x_2, x_3) \oplus \xi_2 \\
c_3 &= \text{parity}(x_1, x_2, x_5) \oplus \xi_3 \\
p &= \text{logistic}(1.0c_1 + 2.0c_2 + 3.0c_3 - 2.0) \\
y &\sim \text{Bernoulli}(p)
\end{aligned}
\tag{B.1}
$$

The distribution in (B.1) includes an explicit noise parameter $p_\xi \in [0,1]$ that we can set to control the noise in concepts. When $p_\xi = 0$, the values of c_1, c_2, c_3 operate as parity functions, which can only be learned through a sufficiently complex model. When $p_\xi > 0$, we inject noise into the concept labels by randomly flipping the values of c_1, c_2, c_3 with probability p_ξ. Thus, the noise parameter sets an upper bound on the accuracy of concept labels – and larger values of p_ξ lead to less accurate concept detectors. In contrast to the real-world datasets, we train the concept detectors for the `noisyconcepts` datasets directly (i.e., without an embedding layer) by fitting a multi-layer perceptron with a single hidden layer.

B.2.2 Additional Experimental Results

Table B.1. Coverage across abstention thresholds τ with confirmation budget 10%

Dataset	Prediction Thresholds	X->Y MLP	Baseline	CS	Baseline + RandomConf	CS + ImpactConf
`warbler`	$\tau = 0.05$	86.00%	17.3%	74.0%	26.0%	85.3%
$n = 5,994$	$\tau = 0.1$	100.00%	74.0%	87.3%	86.7%	100.00%
$m = 112$	$\tau = 0.15$	100.00%	100.00%	100.00%	97.3%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	100.00%	100.00%	100.00%	100.00%
`flycatcher`	$\tau = 0.05$	0.0%	26.9%	0.0%	38.5%	100.00%
$n = 5,994$	$\tau = 0.1$	82.7%	44.2%	36.5%	38.5%	100.00%
$m = 112$	$\tau = 0.15$	92.3%	44.2%	36.5%	38.5%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	57.7%	51.9%	96.2%	100.00%
`melanoma`	$\tau = 0.05$	0.0%	0.0%	4.3%	4.3%	38.57%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	14.3%	4.3%	38.57%
$m = 17$	$\tau = 0.15$	0.0%	0.0%	14.3%	21.4%	100.00%
Kawahara et al. [2018]	$\tau = 0.2$	0.0%	18.6%	34.3%	21.4%	100.00%
`skincancer`	$\tau = 0.05$	0.0%	0.0%	0.0%	3.66%	0.0%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	36.6%	13.4%	37.80%
$m = 17$	$\tau = 0.15$	32.9%	11.0%	36.6%	13.4%	37.80%
Kawahara et al. [2018]	$\tau = 0.2$	86.6%	11.0%	36.6%	30.5%	93.90%
`noisyconcepts25`	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 100,000$	$\tau = 0.1$	32.3%	32.3%	32.3%	33.2%	39.77%
$m = 3$	$\tau = 0.15$	44.1%	43.6%	43.6%	44.8%	65.35%
	$\tau = 0.2$	80.4%	80.4%	80.4%	84.00%	82.7%
`noisyconcepts75`	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 100,000$	$\tau = 0.1$	0.0%	0.0%	0.0%	0.0%	0.0%
$m = 3$	$\tau = 0.15$	0.0%	0.0%	0.0%	0.0%	2.38%
	$\tau = 0.2$	0.0%	0.0%	0.0%	0.0%	2.38%

Table B.2. Coverage across abstention thresholds τ with confirmation budget 50%

Dataset	Prediction Thresholds	X->Y MLP	Baseline	CS	Baseline + RandomConf	CS + ImpactConf
warbler	$\tau = 0.05$	86.0%	17.3%	74.0%	84.7%	86.67%
$n = 5,994$	$\tau = 0.1$	100.00%	74.0%	87.3%	93.3%	92.7%
$m = 112$	$\tau = 0.15$	100.00%	100.00%	100.00%	100.00%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	100.00%	100.00%	100.00%	100.00%
flycatcher	$\tau = 0.05$	0.0%	26.9%	0.0%	67.3%	100.00%
$n = 5,994$	$\tau = 0.1$	82.7%	44.2%	36.5%	76.9%	100.00%
$m = 112$	$\tau = 0.15$	92.3%	44.2%	36.5%	90.4%	100.00%
Wah et al. [2011]	$\tau = 0.2$	100.00%	57.7%	51.9%	100.00%	100.00%
melanoma	$\tau = 0.05$	0.0%	0.0%	4.3%	18.6%	70.00%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	14.3%	30.0%	82.86%
$m = 17$	$\tau = 0.15$	0.0%	0.0%	14.3%	51.4%	88.57%
Kawahara et al. [2018]	$\tau = 0.2$	0.0%	18.6%	34.3%	75.7%	100.00%
skincancer	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 616$	$\tau = 0.1$	0.0%	0.0%	36.6%	28.0%	51.22%
$m = 17$	$\tau = 0.15$	32.9%	11.0%	36.6%	56.1%	100.00%
Kawahara et al. [2018]	$\tau = 0.2$	86.6%	11.0%	36.6%	84.1%	100.00%
noisyconcepts25	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	19.75%
$n = 100,000$	$\tau = 0.1$	32.3%	32.3%	32.3%	36.0%	38.69%
$m = 3$	$\tau = 0.15$	44.1%	43.6%	43.6%	65.6%	78.19%
	$\tau = 0.2$	80.4%	80.4%	80.4%	91.6%	100.00%
noisyconcepts75	$\tau = 0.05$	0.0%	0.0%	0.0%	0.0%	0.0%
$n = 100,000$	$\tau = 0.1$	0.0%	0.0%	0.0%	0.0%	29.55%
$m = 3$	$\tau = 0.15$	0.0%	0.0%	0.0%	0.0%	29.55%
	$\tau = 0.2$	0.0%	0.0%	0.0%	0.0%	68.25%

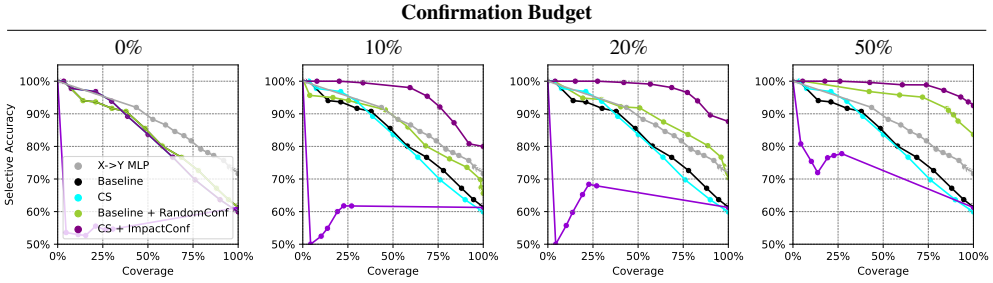
B.2.3 Multiclass Classification Tasks

In this Appendix, we briefly describe how to build conceptual safeguards for multiclass classification tasks and present experimental results for this setting.

In practice, the main requirement for adapting uncertainty propagation is to handle multiple labels. This requires replacing the concept prediction vector with a matrix that encodes $\Pr(y|\boldsymbol{c})$ for all $y \in \mathscr{Y}$ and $\boldsymbol{c} \in \{0,1\}^m$. For the selective classifier φ_τ, we estimate uncertainty based on the likelihood of the most probable class and threshold prediction accordingly.

Table B.3. Coverage vs. accuracy for all methods on multiclass classification tasks.

`cubtypes`
$n = 5,994$
$m = 112$
Wah et al. [2011]



Appendix C

Sufficient Context

C.1 Supporting Experiment Information

We provide sufficient details to reproduce all of our experiments. We list all of the prompts that we use in the next section.

C.1.1 Models

GPT 4o. We use the API-accessible gpt-4o-2024-08-06 model, released on August 6, 2024 [Achiam et al., 2023].

Gemini 1.5 Pro. We use the API-accessible gemini-1.5-pro-0514 model, released on May 14, 2024 [Gemini Team et al., 2023].

Claude 3.5 Sonnet. We use the only API-accessible claude-3-5-sonnet-20240620 model, released on June 20, 2024 [Anthropic, 2024].

Gemma 2 27B. We use the publicly available instruction tuned gemma-2-27b-it model, released on Jun 27, 2024 [Gemma Team et al., 2024].

Mistral 3 7B. We use the publicly available instruction tuned Mistral-7B-Instruct-v0.3 model, released on May 22, 2024 [Jiang et al., 2023].

FLAMe. We use the published FLAMe-RM-24B model [Vu et al., 2024].

TRUE-NLI model. Calculated the maximum probability over the chunks in the context. Use a threshold of 0.05, where if the maximum probability is higher than this, then we classify it

as 'sufficient context'. The threshold of 0.05 achieved the highest F1 score on our human-labeled dataset. We use the `t5_xxl_true_nli_mixture` version of their model [Honovich et al., 2022].

C.1.2 Fine-tuning Settings

In our fine-tuning setup, we employed the LoRA adaptation technique [Hu et al., 2022] to fine-tune Mistral-7B-Instruct-v0.3[1]. We used either a 2,000-example random subset sampled from the training set of the Musique-Ans dataset or from the development set of the HotPotQA data[2]. The prompt template for finetuning is provided in Appendix C.4.

For the LoRA parameters, we set the rank to 4 and alpha to 8 for all experiments. The models were fine-tuned over 2 epochs with a batch size of 16 and a learning rate of 1×10^{-5}. We note that the training was not smooth, and different checkpoints led to very different results. To be systematic, we chose the best checkpoint in terms of Correct % after either 1 or 2 epochs (where for Musique it turned out to be after 1 epoch, and for HotPotQA we found that 2 epochs was better).

C.1.3 Datasets

We sample 500 examples from HotPotQA and Musique-Ans dev sets, following prior work. We use all 'True Premise' questions from FreshQA.

Retrieval for HotpotQA.

We adopt the FlashRAG framework [Jin et al., 2024] to implement our Retrieval-Augmented Generation (RAG) process. Our retrieval corpus is based on the wiki-18 dataset, utilizing 'intfloat/e5-base-v2' from Hugging Face's model hub as a Dense Retriever [3]. For each query, we retrieved the top 5 documents, which were subsequently concatenated with the query and placed within a prompt template for inference.

[1] Available at huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[2] We use the dev set for HotPotQA since the training set had a much different distribution of sufficient context examples. Namely, we found the train set to be over 88% sufficient context, while the dev set was only 44%.

[3] `huggingface.co/intfloat/e5-base-v2`

To explore advanced retrieval techniques, we also evaluated the REPLUG [Shi et al., 2023b] method. REPLUG enhances the generation quality by prepending each retrieved document individually to the input context and ensembling output probabilities across different passes. The REPLUG method is also implemented based on the FlashRAG framework [Jin et al., 2024].

Retrieval for FreshQA

We use the URLs provided in the FreshQA dataset as retrieval for the `context`. We scraped each URL and discarded extra HTML content such as headers, footers, and navigation. We include the title of each webpage in the text, convert any included tables to markdown, and include the table title immediately before the table in the text. When splitting the tables and text for smaller context windows, we keep tables and sentences intact when possible. For large tables that require splitting, we duplicate the table row column headers to include them in each chunk.

C.2 Additional Results

C.2.1 Fine-tuning Full Results

One aspect of our selective generation framework is that we use FLAMe, a 24B model, to provide sufficient context labels. However, we would incur significant overhead if we used a 24B model to improve the generation of a much smaller LLM. Instead, we try directly fine-tuning Mistral 3 7B to increase accuracy with retrieval. Specifically, we experiment with different data mixtures to encourage the model to output "I don't know" instead of generating an incorrect response.

Fine-tuning Data. We repeat the following process separately for each of the Musique-Ans and HotPotQA datasets to create three mixtures of training data with different answers for each dataset. First, we sample 2000 instances. Then, for Data Mix 1, we fine-tune on these instances and keep their given ground truth answer. For Data Mix 2, we choose 400 examples (20%) at random and change the answer to "I don't know" before fine-tuning. For Data Mix 3, we instead randomly choose 400 examples (20%) that our autorater labels as insufficient context and change their answer to "I don't know" while keeping the other answers as the ground truth. We hypothesize that fine-tuning on Data Mix 2 and 3 should steer the model to abstain more and hallucinate less than with Data Mix 1.

Models, Methods, Metrics. Using the three data mixtures described above, we fine-tune the Mistral 3 7B Instruct model model with LoRA (details in Appendix C.1.2). At inference time, we use the standard RAG setup where we add context to the prompt. As baselines, we also evaluate the model without fine-tuning in both the closed-book setting (w/o RAG) and the open-book setting (Vanilla RAG). Consistent with the prior experiments, we use an LLM with ground truth answers to classify responses as Correct (**%C**), Abstention (**%A**), or Hallucination (**%H**).

Fine-tuning Results and Discussion. Table C.1 shows our experimental results. We verify that the FT variants have a higher rate of generating correct answers (**%C**) compared to

closed-book and Vanilla RAG for Musique, but not for HotPotQA. On the other hand, refuting our hypothesis, Data Mix 2 and 3 do not lead to more abstentions than Vanilla RAG. But, they do abstain more than with Data Mix 1, showing the impact of adding "I don't know" in the training set. In general, FT models using RAG output incorrect answers (**%H**) much of the time, and often more than they abstain (**%A**).

Table C.1. Fine-tuned (FT) Mistral 3 7B Instruct. We compare closed book and vanilla RAG with three FT settings, measuring % Correct (**%C**), % Abstain (%A), and % Hallucinate (**%H**). Also, "idk" means we change the answer in training samples to be "I don't know" instead of the given answer (either for 20% of random examples, or 20% of examples with insufficient context). Best **%C** for each model/dataset in bold.

			Musique			HotPotQA		
Model	**Variant**	RAG	**%C**	**%A**	**%H**	**%C**	**%A**	**%H**
Mistral	Closed Book		6.6	29.8	63.6	32	7.6	60.4
"	Vanilla RAG	✓	28.8	11.8	59.4	**46.6**	9.2	44.2
"	FT GT answer (Data Mix 1)	✓	**31.4**	0	68.6	43.4	0	56.6
"	FT idk 20% rand. (Data Mix 2)	✓	23	1.2	75.8	41.6	0.8	57.6
"	FT idk 20% insuff. (Data Mix 3)	✓	23	2.2	74.8	41.2	2	56.8

C.2.2 Performance Breakdown by Sufficient Context

We explore RAG performance by different models for various RAG benchmark datasets. Here, the first column shows performance without RAG (closed-book) while the second column shows performance with RAG (open-book). To better understand RAG performance, we use our sufficient context autorater to stratify the retrieval augmented generation (RAG) datasets into sufficient and insufficient context. The third and fourth columns show the performance of the second column stratified by sufficient vs insufficient context, respectively.

Table C.2. Performance Analysis of RAG Systems Using Human-Annotated Sufficient Context Labels. These tables include results on a curated set of challenging context-dependent questions. Table (a) shows that while larger models generally achieve higher accuracy with sufficient context (present in 54.8% of cases), even top performers exhibit a 14-16% error rate. Table (b) reveals that with insufficient context (45.2% of cases), models predominantly abstain from answering (50-73% of instances), though significant hallucination rates (15-40%) persist. These patterns of context-dependent performance and hallucination risk are consistent with our analyses of HotpotQA, FreshQA, and Musique datasets, despite variations in absolute performance due to different task complexities.

(a) Performance with Sufficient Context (54.8% of Dataset)

Model	% Correct	% Abstain	% Hallucinate
Gemini 1.5 Pro	84.1	1.6	14.3
GPT 4o	82.5	4.8	12.7
Claude 3.5 Sonnet	85.7	11.1	3.2
Gemini 1.5 Flash	77.8	4.8	17.5
Gemma 27B	71.4	3.2	25.4

(b) Performance with Insufficient Context (45.2% of Dataset)

Model	% Correct	% Abstain	% Hallucinate
Gemini 1.5 Pro	9.6	50.0	40.4
GPT 4o	23.1	61.5	15.4
Claude 3.5 Sonnet	9.6	53.8	36.5
Gemini 1.5 Flash	7.7	73.1	19.2
Gemma 27B	9.6	55.8	34.6

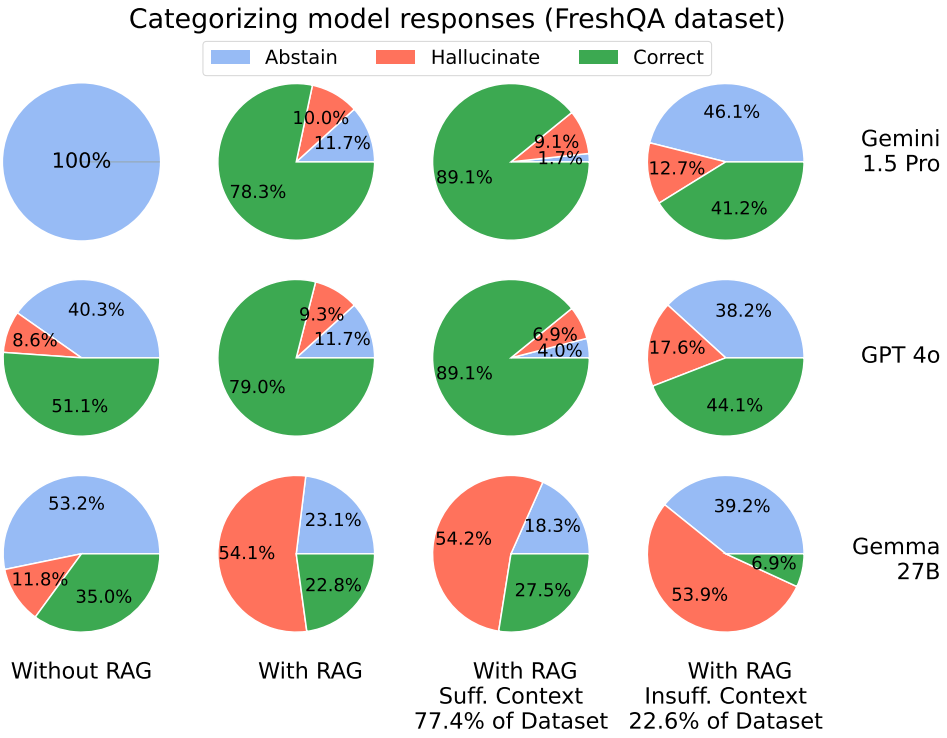


Figure C.1. Correct, hallucination, and abstention fractions across models for dataset FreshQA, stratified by sufficient context. FreshQA includes hand-curated source URLs, which explains the larger percentage of sufficient context (77.4%). FreshQA also specifically explores questions with answers that change based on the question's timestamp, which may explain the frequent abstentions without RAG (100% for Gemini 1.5 Pro).

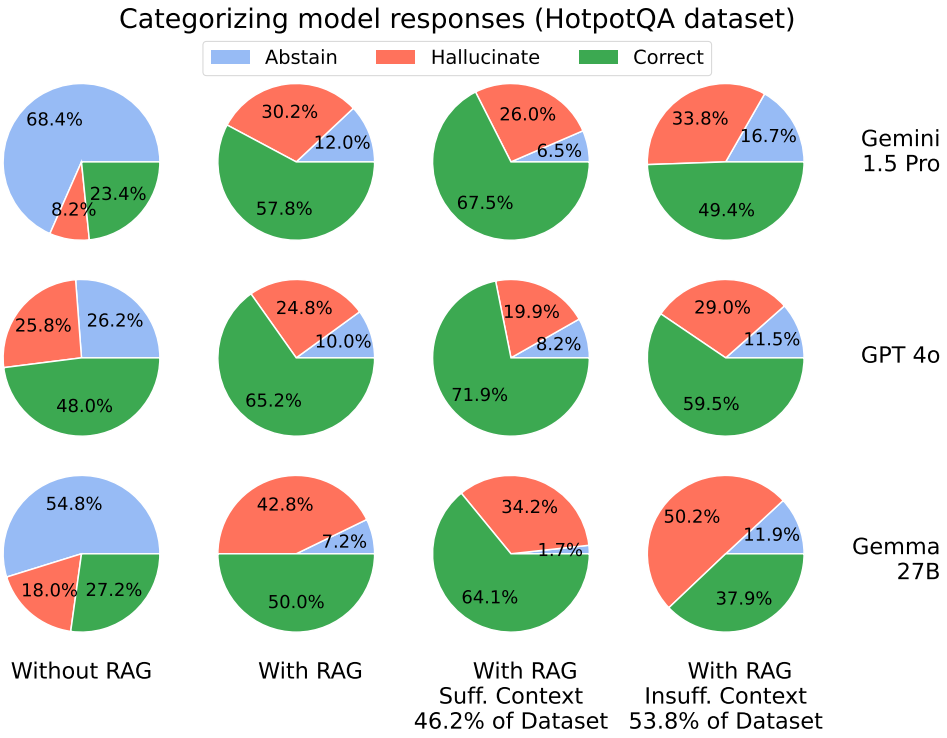


Figure C.2. Correct, hallucination, and abstention fractions across models for dataset HotpotQA, stratified by sufficient context. HotpotQA includes questions that are more likely to be answerable without context (e.g., yes or no questions, multiple choice questions, or questions with answers that might be answerable due to pre-training). This explains the higher fraction of correct answers without RAG (e.g., 48.0% for GPT 4o).

C.2.3 Comparison of QA Evaluation Metrics

We compare two LLM-based QA Evaluator (LLMEval) used in the paper with a deterministic lexical matching metric (Contains Answer). The Contains Answer metric labels responses based on whether they contain the exact ground truth answer, while LLMEval uses an LLM to assess semantic correctness.

Table C.3 presents model performance across three datasets (FreshQA, Musique, HotpotQA), split by our sufficient context autorater. The results show that Contains Answer is generally stricter than LLMEval, though both metrics reveal similar patterns in model behavior.

Table C.3. Comparison of evaluation metrics across models and datasets. We show results for checking whether the response contains one of the ground truth answer strings ("Contains"), where we report the % of responses that contain an answer. We compare this to our LLMEval method that uses an LLM to evaluate if the response is correct, abstain, or hallucinated, and we report % correct.

		FreshQA		Musique		HotpotQA	
Model	Context	Contains	LlmEval	Contains	LlmEval	Contains	LlmEval
Gemini 1.5	Suff	80.3%	89.1%	60.1%	83.4%	47.6%	67.5%
Pro	Insuff	31.4%	41.2%	33.6%	49.5%	34.2%	49.4%
GPT-4	Suff	84.3%	89.1%	64.6%	83.4%	52.4%	71.9%
	Insuff	36.3%	44.1%	44.4%	61.4%	46.1%	59.5%
Gemma	Suff	26.9%	27.5%	10.8%	23.3%	40.7%	64.1%
	Insuff	11.8%	6.9%	7.2%	10.1%	22.7%	37.9%
Claude 3.5	Suff	67.9%	73.1%	48.9%	74.0%	46.3%	66.7%
Sonnet	Insuff	26.5%	33.3%	19.9%	40.4%	29.0%	38.3%

The Contains Answer metric exhibits several characteristics when compared to LLMEval:

1. Different formatting affects matching:

```
Q: What date did the creator of Autumn Leaves die?
Ground Truth: 13 August 1896
Response: August 13, 1896.
Contains Answer: False
LLMEval: Correct
```

2. Semantic equivalents are not captured:

```
Q: What former Los Angeles Lakers majority owner is the
father of Jeanie Marie Buss?
Ground Truth: Gerald Hatten Buss
Response: Jerry Buss.
Contains Answer: False
LLMEval: Correct
```

3. Partial matches can be marked as correct:

```
Q: What is Amazon Prime Video's most-watched premiere ever?
Ground Truth: The Rings of Power
Response: The series explores the Rings of Power
the rise of Sauron...
Contains Answer: True
LLMEval: Hallucinate
```

The LLM QA evaluator provides several practical advantages:

- Handles variations in model verbosity and formatting
- Distinguishes between correct, abstain, and incorrect responses
- Enables efficient evaluation across multiple datasets

Our analysis shows two key findings that are consistent across both metrics: LLMs (i) exhibit hallucination even with sufficient context and (ii) struggle to abstain with insufficient context.

C.3 Prompts

C.3.1 Sufficient Context Autorater Prompt

You are an expert LLM evaluator that excels at evaluating a QUESTION and REFERENCES. Consider the following criteria:

Sufficient Context: 1 IF the CONTEXT is sufficient to infer the answer to the question and 0 IF the CONTEXT cannot be used to infer the answer to the question

Assume the queries have timestamp <TIMESTAMP>.

First, output a list of step-by-step questions that would be used to arrive at a label for the criteria. Make sure to include questions about assumptions implicit in the QUESTION. Include questions about any mathematical calculations or arithmetic that would be required.

Next, answer each of the questions. Make sure to work step by step through any required mathematical calculations or arithmetic. Finally, use these answers to evaluate the criteria.

Output the ### EXPLANATION (Text). Then, use the EXPLANATION to output the ### EVALUATION (JSON)

EXAMPLE:

QUESTION

In which year did the publisher of Roald Dahl's Guide to Railway Safety cease to exist?

References

Roald Dahl's Guide to Railway Safety was published in 1991 by the British Railways Board. The British Railways Board had asked Roald Dahl to write the text of the booklet, and Quentin Blake to illustrate it, to help young people enjoy using the railways safely. The British Railways Board (BRB) was a nationalised industry in the United Kingdom that operated from 1963 to 2001. Until 1997 it was responsible for most railway services in Great Britain, trading under the brand name British Railways and, from 1965, British Rail. It did not operate railways in Northern Ireland, where railways were the responsibility of the Government of Northern Ireland.

EXPLANATION

The context mentions that Roald Dahl's Guide to Railway Safety was published by the British Railways Board. It also states that the British Railways Board operated from 1963 to 2001, meaning the year it ceased to exist was 2001. Therefore, the context does provide a precise answer to the question.

JSON

{"Sufficient Context": 1}

Remember the instructions: You are an expert LLM evaluator that excels at evaluating a QUESTION and REFERENCES. Consider the following criteria:

Sufficient Context: 1 IF the CONTEXT is sufficient to infer the answer to the question and 0 IF the CONTEXT cannot be used to infer the answer to the question

Assume the queries have timestamp TIMESTAMP.

First, output a list of step-by-step questions that would be used to arrive at a label for the criteria. Make sure to include questions about assumptions implicit in the QUESTION Include questions about any mathematical calculations or arithmetic that would be required.

Next, answer each of the questions. Make sure to work step by step through any required mathematical calculations or arithmetic. Finally, use these answers to evaluate the criteria. Output the ### EXPLANATION (Text). Then, use the EXPLANATION to output the ### EVALUATION (JSON)

QUESTION

<question>

REFERENCES

<context>

C.3.2 FLAMe Prompt

INSTRUCTIONS:

title: Is the context sufficient to infer the answer to the question?

description: In this task, you will be provided with documents and a question. Use one of the following labels under 'judgment':

1. sufficient: The documents are not sufficient to infer the answer to the question.

2. insufficient: The documents are sufficient to infer the answer to the question.

output_fields: judgment

CONTEXT:

documents:<references> question: <question>

C.3.3 LLMEval Prompt

Since the questions in our datasets ask for free-form answers, the LLM responses may not exactly match the GT answers. Hence, we use an LLM to determine: the answers are the same (Correct), or the LLM does not answer the question (Abstain), or the answer is incorrect (Hallucinate). We note that prior work has shown that Gemini 1.5 Pro has very high accuracy and correlation with human judgments for evaluating free-form responses [Krishna et al., 2024]. Responses resulting in empty strings were classified as "missing," while variations of "I don't know" were also treated as missing. We normalized both the ground truth answers and the model's responses by removing punctuation, converting to lowercase, and eliminating stop words.

===Task===

I need your help in evaluating an answer provided by an LLM against ground truth answers. Your task is to determine if the LLM's response matches the ground truth answers. Please analyze the provided data and make a decision.

===Instructions===

1. Carefully compare the "Predicted Answer" with the "Ground Truth Answers". 2. Consider the substance of the answers – look for equivalent information or correct answers. Do not focus on

exact wording unless the exact wording is crucial to the meaning.

3. Your final decision should be based on whether the meaning and the vital facts of the "Ground Truth Answers" are present in the "Predicted Answer." 4. Categorize the answer as one of the following:

- "perfect": The answer is completely correct and matches the ground truth.
- "acceptable": The answer is partially correct or contains the main idea of the ground truth.
- "incorrect": The answer is wrong or contradicts the ground truth.
- "missing": The answer is "I don't know", "invalid question", or similar responses indicating lack of knowledge.

===Input Data===

- Question: What 1876 battle featured the Other Magpie?
- Predicted Answer: The Other Magpie fought in the Battle of the Rosebud.
- Ground Truth Answers: Battle of the Rosebud

===Output Format===

Provide your evaluation in the following format:

Explanation: (How you made the decision)

Decision: (One of "perfect", "acceptable", "incorrect", or "missing")

Please proceed with the evaluation.

C.3.4 Dataset Question Answer Prompts

The CoT prompt instructs the model to provide an accurate and concise answer based solely on the given search results, using an unbiased and journalistic tone. The prompt includes an example question, references, and an answer to guide the model's response format. To extract the final answer, we implemented a pattern-matching technique on the model's response,

specifically targeting the text following "The answer is:" for CoT prompts.

Chain of Thought (CoT) Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant). Start with an accurate, engaging, and concise explanation based only on the provided documents. Must end with "The answer is:". Use an unbiased and journalistic tone.

EXAMPLE:

Question

<example question>

References

<example references>

Answer

<example answer>

Question

<question>

References

<references>

Answer

Answer Only (AO) Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant). Do not say anything other than the answer itself.

EXAMPLE:

Question

<example question>

References

<example references>

Answer

<example answer>

Question

<question>

References

<references>

Answer

C.4 Fine-Tuning and RAG Prompts for Mistral

Finetuning Prompt (FT) Answer the question based on the given document. Only give me the answer and do not output any other words. The following are given references.

References

<references>

Please follow the following guideline when formulating your answer: if you are uncertain or don't know the answer, respond with "I don't know".

Question

<question>

Answer

<answer>

Evaluation Without RAG Prompt Answer the question based on your own knowledge. Only give me the answer and do not output any other words. Please follow the following guideline when formulating your answer: if you are uncertain or don't know the answer, respond with "I don't know".

Question

<question>

Answer

Evaluation With RAG Prompt Answer the question based on the given document. Only give me the answer and do not output any other words. The following are given references.

References

<references>

Please follow the following guideline when formulating your answer: if you are uncertain or don't know the answer, respond with "I don't know".

Question

<question>

Answer

Bibliography

Aaron Abajian, Nikitha Murali, Lynn Jeanette Savic, Fabian Max Laage-Gaupp, Nariman Nezami, James S Duncan, Todd Schlachter, MingDe Lin, Jean-Francois Geschwind, and Julius Chapiro. Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning—an artificial intelligence concept. *Journal of Vascular and Interventional Radiology*, 29(6):850–857, 2018.

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.

Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

Jérôme Allyn, Cyril Ferdynus, Michel Bohrer, Cécile Dalban, Dorothée Valance, and Nicolas Allou. Simplified acute physiology score ii as predictor of mortality in intensive care units: a decision curve analysis. *PloS one*, 11(10):e0164828, 2016.

Catherine L Anderson and Ritu Agarwal. The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information. *Information Systems Research*, 22(3):469–490, 2011.

Anthropic. Claude 3.5 sonnet model card addendum, 2024.

April Moreno Arellano, Wenrui Dai, Shuang Wang, Xiaoqian Jiang, and Lucila Ohno-Machado. Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1:115, 2018.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*, 2024.

Onur Atan, William Whoiles, and Mihaela Schaar. Data-driven online decision making with costly information acquisition. *Arxiv*, 02 2016.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center: Internet, Science and Tech*, 2019.

Naveen Farag Awad and Mayuram S Krishnan. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, pages 13–28, 2006.

Gaurav Bansal, David Gefen, et al. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision support systems*, 49(2):138–150, 2010.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 399–408, 2020.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Max Biggs, Wei Sun, and Markus Ettl. Model distillation for revenue optimization: Interpretable personalized pricing. *arXiv preprint arXiv:2007.01903*, 2020.

P. Bukaty. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, 2019. ISBN 9781787781337. URL https://books.google.com/books?id=vGWfDwAAQBAJ.

Tim S Campbell and William A Kracaw. Information production, market signalling, and the theory of financial intermediation. *the Journal of Finance*, 35(4):863–882, 1980.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine*

Learning, pages 1507–1517. PMLR, 2021.

Thomas J Chemmanur. The pricing of initial public offerings: A dynamic model with information production. *The Journal of Finance*, 48(1):285–304, 1993.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications*, 11(1):6386, 2020.

C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, pages 247–254, 1957.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*, 2024.

Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. Human uncertainty in concept-based ai systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 869–889, 2023.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, 2024.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, pages 189–212, 1996.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Online certification

of preference-based fairness for personalized recommender systems. *arXiv preprint arXiv:2104.14527*, 2021.

Adrian GK Edwards, Gurudutt Naik, Harry Ahmed, Glyn J Elwyn, Timothy Pickles, Kerry Hood, and Rebecca Playle. Personalised risk communication for informed decision making about taking screening tests. *Cochrane database of systematic reviews*, Cochrane database of systematic reviews(2), 2013.

Kfir Eliaz and Ran Spiegler. On incentive-compatible estimators. *Games and Economic Behavior*, 132:204–220, 2022.

Adam N Elmachtoub, Vishal Gupta, and Michael Hamilton. The value of personalized pricing. *Available at SSRN 3127719*, 2018.

European Parliament and of the Council. Regulation 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Official Journal of the European Union.

Haiyan Fan and Marshall Scott Poole. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202, 2006.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, 2024.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*, 2022.

Vaclav Voracek Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Giorgio Fumera and Fabio Roli. Reject option with multiple thresholds. *Pattern recognition*, 33 (12):2099–2101, 2000.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, 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.

Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun 2022. doi: 10.1145/3531146.3533172. URL http://dx.doi.org/10.1145/3531146.3533172.

Jason S Haukoos, Michael S Lyons, Christopher J Lindsell, Emily Hopkins, Brooke Bender, Richard E Rothman, Yu-Hsiang Hsieh, Lynsay A MacLaren, Mark W Thrun, Comilla Sasson, et al. Derivation and validation of the denver human immunodeficiency virus (hiv) risk score for targeted hiv screening. *American journal of epidemiology*, 175(8):838–846, 2012.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, pages 1944–1953, 2018.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

SM Hollenberg. Cardiogenic shock. In *Intensive Care Medicine*, pages 447–458. Springer, 2003.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, 2022.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 14982–14995, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.890. URL https://aclanthology.org/2024.findings-acl.890.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*, 2023.

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. *arXiv preprint arXiv:2407.13101*, 2024.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576, 2024. URL https://arxiv.org/abs/2405.13576.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Hailey Joren, Chirag Nagpal, Katherine A Heller, and Berk Ustun. Participatory personalization in classification. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2dbb8bfe4cd3875609b23799830ee865-Paper-Conference.pdf.

Hailey Joren, Charles Marx, and Berk Ustun. Classification with conceptual safeguards. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://iclr.cc/virtual/2024/poster/17625.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Margot E Kaminski. The right to explanation, explained. *Berkeley Tech. LJ*, 34:189, 2019.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerryman-

dering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

Ronald C Kessler, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, EVA Hiripi, Mary J Howes, Robert Jin, Kristina Secnik, Thomas Spencer, et al. The world health organization adult adhd self-report scale (asrs): a short screening scale for use in the general population. *Psychological medicine*, 35(2):245–256, 2005.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

Michael P Kim, Aleksandra Korolova, Guy N Rothblum, and Gal Yona. Preference-informed fairness. *arXiv preprint arXiv:1904.01793*, 2019.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*, 2024.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318, 2021.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24): 2957–2963, 1993.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *arXiv preprint arXiv:2403.09972*, 2024.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J Salganik. Privacy, ethics, and data access: A case study of the fragile families challenge. *Socius*, 5:2378023118813023, 2019.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Paul Mineiro. Online joint fine-tuning of multi-agent flows. *arXiv preprint arXiv:2406.04516*, 2024.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In *COLM 2024*, 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585.

OECD. Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data, 2013. URL https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Martin Ortlieb and Ryan Garner. Sensitivity of personal data items in different online contexts. *it-Information Technology*, 58(5):217–228, 2016.

Lucas Monteiro Paes, Carol Xuan Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Snp3iEj7NJ.

Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1):

1–8, 2020.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Vassilina Nikoulina, and Stéphane Clinchant. Bergen: A benchmarking library for retrieval-augmented generation. *arXiv preprint arXiv:2407.01102*, 2024.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*, 2024.

Emil Scosyrev, James Messing, Katia Noyes, Peter Veazie, and Edward Messing. Surveillance epidemiology and end results (seer) program and population-based research in urologic oncology: an overview. In *Urologic Oncology: Seminars and Original Investigations*, volume 30, pages 126–132. Elsevier, 2012.

Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 839–849, 2022.

Rachel Sharp. The hamilton rating scale for depression. *Occupational Medicine*, 65(4):340–340, 2015.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023a.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models.

arXiv preprint arXiv:2301.12652, 2023b.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Constructing benchmarks and interventions for combating hallucinations in llms. *arXiv preprint arXiv:2404.09971*, 2024.

Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019.

Aaron F Struck, Berk Ustun, Andres Rodriguez Ruiz, Jong Woo Lee, Suzette M LaRoche, Lawrence J Hirsch, Emily J Gilmore, Jan Vlachy, Hiba Arif Haider, and Cynthia Rudin. Association of an electroencephalography-based risk score with seizure probability in hospitalized patients. *JAMA neurology*, 74(12):1419–1424, 2017.

Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21 (11):1389–1393, 2014. doi: 10.1109/LSP.2014.2337313.

Vinith M Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

Cuong Tran and Ferdinando Fioretto. Personalized privacy auditing and optimization at test time. *arXiv preprint arXiv:2302.00077*, 2023.

Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34: 27555–27565, 2021.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Is multihop qa in dire condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, 2020.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for*

Computational Linguistics, 10:539–554, 2022.

U.S. Congress. Health insurance portability and accountability act of 1996, 1996. URL https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Public Law 104-191.

Berk Ustun, M Brandon Westover, Cynthia Rudin, and Matt T Bianchi. Clinical prediction models for sleep apnea: the importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(02):161–168, 2016.

Berk Ustun, Lenard A Adler, Cynthia Rudin, Stephen V Faraone, Thomas J Spencer, Patricia Berglund, Michael J Gruber, and Ronald C Kessler. The world health organization adult attention-deficit/hyperactivity disorder self-report screening scale for dsm-5. *Jama psychiatry*, 74(5):520–526, 2017.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382, 2019.

Gregory Vaughan, Robert Aseltine, Kun Chen, and Jun Yan. Efficient interaction selection for clustered data via stagewise generalized estimating equations. *Statistics in Medicine*, 39(22): 2855–2868, 2020.

Davide Viviano and Jelena Bradic. Fair policy targeting. *arXiv preprint arXiv:2005.12395*, 2020.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*, 2024.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*, 2024a.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, et al. Speculative rag: Enhancing

retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*, 2024b.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. How easily do irrelevant inputs skew the responses of large language models? In *COLM 2024*, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024.

An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*, 2024.

Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R. Bharat Rao. Active sensing. In *AISTATS*, 2009.

Tan Yu, Anbang Xu, and Rama Akkiraju. In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*, 2024.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 228–238, 2017.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132, 2024a.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. End-to-end beam retrieval for multi-hop question answering. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1718–1731, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.96. URL https://aclanthology.org/2024.naacl-long.96.

Wenying Zhou, Yang Yang, Cheng Yu, Juxian Liu, Xingxing Duan, Zongjie Weng, Dan Chen, Qianhong Liang, Qin Fang, Jiaojiao Zhou, et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. *Nature communications*, 12(1):1259, 2021.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1453–1463, 2024.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.