UC Davis
UC Davis Electronic Theses and Dissertations

Title

Trust and Transparency in Deep Learning, from Medical Imaging to Language Models

Permalink

https://escholarship.org/uc/item/3022b9b8

ISBN

9798193262190

Author

Broecker, Stefan

Publication Date

2026-09-08

Peer reviewed|Thesis/dissertation

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

Trust and Transparency in Deep Learning, from Medical Imaging to Language Models

By

STEFAN BROECKER

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Thomas Strohmer, Chair

Jason Yeates Adams

Raissa D'Souza

Committee in Charge

2026

© Stefan Broecker, 2026. All rights reserved.

Contents

Abstract v

Acknowledgments vi

Chapter 1. Introduction 1

Chapter 2. Multimodal Deep Learning for ARDS Detection 4

2.1. Abstract 4

2.2. Introduction 5

2.3. Methods 7

2.4. Results 14

2.5. Discussion 17

2.6. Conclusion 24

2.7. Acknowledgments 24

2.8. Ethics Statement 24

Chapter 3. Optimal Transport Learning through Synchronized Forward-Backward Interpolants 26

3.1. Abstract 26

3.2. Introduction 27

3.3. Background and Problem Setup 32

3.4. Proposed Framework 36

3.5. Theoretical Motivation Underlying the Forward-Backward Supervision 47

3.6. Experiments 56

3.7. Conclusion 70

Chapter 4. The “I Don’t Know” Filter: Enhancing Agentic Reliability in Function Calling 72

4.1. Abstract 72
4.2. Introduction 73
4.3. Background 74
4.4. Methods 76
4.5. Results 81
4.6. Discussion 83
4.7. Conclusions 86
Acknowledgement 87

Chapter 5. Forgetting-MarI: LLM Unlearning via Marginal Information Regularization 88
5.1. Abstract 88
5.2. Introduction 89
5.3. Unlearning via Marginal Information 93
5.4. Algorithm Design 98
5.5. Experiments 99
5.6. Conclusion 106
5.7. Broader Impacts 106

Chapter 6. Estimating Temperature in LLM-Generated Text: A Maximum Likelihood Approach and Its Limits 108
6.1. Abstract 108
6.2. Introduction 109
6.3. Background and Approach 110
6.4. Theoretical Analysis 112
6.5. Simulation-Based Validation 114
6.6. Experiments with Language Models 115
6.7. Discussion 118

Appendix A. Appendix for Chapter 2 122

Appendix B. Appendix for Chapter 3 125
B.1. Proofs 125

B.2. Training Algorithm 126
B.3. Additional Experimental Results 127

Appendix C. Appendix for Chapter 5 129
C.1. Proof of Proposition 5.3.1 129
C.2. Generalization of Theorem 5.3.2 to unlearn set neighborhood 130
C.3. Proof of Theorem 5.3.2 132
C.4. Theoretical error bound between position-wise vs. pooled MarI 133
C.5. Theoretical Guarantee provided by pooled MarI 135

Appendix D. Appendix for Chapter 6 137
D.1. Deriving An Equation for Fisher Information 137
D.2. Full Derivative Calculations 138
D.3. Full Derivative "Calculations" with SymPy 142

Bibliography 144

Abstract

As deep learning models have grown deeper and more complex, and as they have been deployed in high stakes decision-making settings, their lack of transparency has become an increasingly urgent problem. This dissertation addresses that problem across two domains, medical artificial intelligence (AI) and large language models (LLMs), with the shared goal of improving our ability to understand, characterize, and ultimately trust the outputs of black box models.

The first part of this work lies within a broader effort to build AI-driven, patient-specific synthetic data capable of supporting tasks like diagnosis. This dissertation lays the groundwork for that goal in two ways. First, it shows that combining physiologically heterogeneous data sources in a shared latent space can improve the diagnostic performance of a deep learning model on acute respiratory distress syndrome (ARDS). Second, it demonstrates the ability of optimal transport-based models to generate synthetic chest x-rays by selectively and reliably manipulating disease-relevant features of images, and introduces a new, more effective training method to that end. The second part of this dissertation turns to LLMs. First, it introduces a method for improving the reliability of agentic tool use by quantifying the uncertainty of LLM responses and learning when models should say "I don't know." It then presents a mathematically rigorous and verifiable approach to machine unlearning, addressing the growing need to remove information from trained models for legal, regulatory, and data management reasons. Finally, this dissertation explores a novel method for quantifying the temperature used during LLM text generation and the lessons that can be learned from where the method fails.

Together, these contributions span medical imaging and natural language processing, but they share a common throughline that trust in deep learning systems is based on an understanding of how they behave and where their limitations lie.

Acknowledgments

To start with, a huge thank you to my advisor and committee chair, Thomas Strohmer. I genuinely couldn't have asked for a better mentor for my PhD. You gave me the support and freedom to explore academically and professionally, and also knew when I needed to be reined in. I look forward to many more years of collaboration.

Thank you to my other two committee members, Jason Adams and Raissa D'Souza, both for being a part of my committee and for shaping my UC Davis experience. I learned so much about being a good scientist from working with and watching both of you. I can't thank you enough.

Thank you also to Norman Matloff. You'll never get the chance to read this, but I hope you knew how much you positively affected my path at UC Davis and how grateful I am that we were able to work together.

Thank you to all of my collaborators during my time at UC Davis (I know I'm forgetting some people, and I apologize in advance): Shizhou Xu, Girish Kumar, Yuan Ni, Jon Forstater, and Jiachen Liu from UC Davis; Boris Selitser and Mason del Rosario from Okareo; and Anna Liu from UC Davis Health.

Lastly, thank you to my mom and sister, Vera and Katja Broecker. Yes, you have to call me Dr. Broecker now.

CHAPTER 1

Introduction

With the rise of deep learning came the rise of the black box model, especially as deep learning models have gotten deeper and more complex. Further, as these models get put in positions to influence or make decisions in high stakes situations, the consequences of that opacity have become harder to ignore. Namely, a model that we can't explain is a model whose mistakes are hard to catch and whose successes are hard to fully trust. This has led to the development of fields like Explainable Artificial Intelligence (XAI) (Hassija et al., 2024), which try to understand the goings on of deep learning models, as well as calls for people to stop using deep learning entirely for high stakes decisions due to the models' inherent unexplainability (Rudin, 2019). This tension between how powerful deep learning models are and how little we can say about why they work is, in many ways, what drew me to pursue this PhD in the first place, and it has shaped the overarching question that ties this dissertation together: how can we pull the curtain back on black box models, and how can we use the insights we gain from that to improve the trust we have in their outputs?

This dissertation comes to you in two parts. The first is in the realm of medical artificial intelligence (AI), and presents work that is part of a broader effort aimed at making the concept of a digital twin a reality in medicine. A digital twin is a digital replica, often with the aid of AI, that models a system, ranging in scale from organelles to full humans (Erol et al., 2020; Sun et al., 2023). Digital twins can help predict phenomena related to a patient, including making diagnoses. The works presented here set the groundwork for a project where the ultimate goal is to create a digital twin of sorts: the ability to generate multimodal synthetic data. This means data coming from multiple data sources that all correspond to the same patient and thus share the same conditions and features.

The first relevant chapter, Chapter 2, explores the power of joint embeddings in multimodal deep learning models. It does this by looking at how the combination of multimodal data through

joint embeddings affects the diagnosis of a physiologically heterogeneous condition, acute respiratory distress syndrome (ARDS). Joint embeddings are a common method for translating between modalities, and are often trained through a process called contrastive learning, which was popularized in the multimodal setting in works like contrastive language-image pretraining (CLIP) (Chen et al., 2020; Radford et al., 2021). This work explores the idea that creating a joint embedding by embedding modalities in a common latent space can improve the ability of a model to use those modalities for a common goal, which in this case is classification. A key idea of the work is that, by taking in information from more physiologically relevant modalities when making a decision, the model can improve the reliability of its predictions and become a more trustworthy tool to relevant decision makers.

The next relevant chapter, Chapter 3, explores generating synthetic data, specifically chest x-rays. The road to the approach explored in this chapter, optimal transport, includes tests with the some of the most well-known technologies in modern image generation, including diffusion models and generative adversarial networks (GANs). It tackles synthetic x-ray generation through the lens of isolating what in a chest x-ray distinguishes a particular condition and learning to selectively add or remove that condition while keeping the rest of the x-ray constant. The chapter focuses on being rigorous about how we train the model to make that addition or subtraction in order to generate data that medical professionals can trust to be accurate.

As previously mentioned, these two chapters are part of an overall whole that ultimately aims toward multimodal synthetic data generation, where a digital twin is built from several data sources, generated in coordination with one another. The road map for getting there involves extending the joint embedding approach from Chapter 2 beyond classification and using it to coordinate generation across modalities. In so doing, models can generate a synthetic chest x-ray and a synthetic lab value, for instance, that are plausible reflections of the same patient. That coordination is the piece this dissertation does not yet solve, and it remains an open thread for future work.

The second part of my work on pulling back some of the layers of deep learning is in the realm of large language models (LLMs), and shifts the lens from medical data to text. The first chapter, Chapter 4, explores a method for improving our trust in agentic AI. In agentic AI, LLM-powered

agents interact directly with the world around them, often through tools or APIs. This is opposed to language-centric LLMs that interact with users in a conversational style. Of note, since LLM agents perform actions directly, without a human in the loop, a trustworthy agent means trusting that they are performing the action that the user expects them to take. This chapter explores a technique for quantifying the uncertainty of agentic tool use, improving their reliability, and therefore our trust in them.

The next chapter, Chapter 5, deals with an issue that has grown in importance as LLMs have grown in prominence, and that's the ability to remove information from them, or "unlearn" information. Spanning use cases from GDPR's "right to be forgotten" to copyright to simple data management, it's become clear that LLMs need an ability to remove information without fully retraining them from scratch, a potentially extremely costly proposition. This chapter explores a method for unlearning that that is mathematically rigorous and verifiable, providing a method that can be trusted from a technical, social, and legal perspective.

The last chapter, Chapter 6, explores the idea of whether you can infer parameters, specifically temperature, used during the text generation process just based on the text that gets generated. It fits into the broader landscape of model characterization, where researchers try to reverse engineer characteristics of a model based on it's behavior. It also explores the idea that different language models encode language and meaning in different, maybe slightly incompatible, ways. On both levels, the chapter improves our understanding of how language models work, which allows us to make more informed decisions about how much we decide to trust them.

Each of these chapters takes on trust in deep learning from a different angle, covering richer inputs, more careful generation, quantified uncertainty, verifiable removal, and deeper characterization, but they share a common thread: trust in AI is earned through understanding and transparency. The chapters that follow take up each of these threads in turn, moving from medical imaging to language models, all in service of the same goal. That goal is to improve our understanding of the tools we are building, allowing us to make more informed decisions.

CHAPTER 2

Multimodal Deep Learning for ARDS Detection

This chapter is based on a paper currently under review and is a collaboration with Jason Y. Adams and Rachael A. Callcut of UC Davis Health, Girish Kumar and Yuan Ni of the UC Davis Department of Mathematics (at the time of the work), and Thomas Strohmer of the UC Davis Department of Mathematics. This work trains a multimodal deep learning model, using joint embeddings to combine data collected from electronic health records, chest x-rays, and continuously collected ventilation waveforms, to diagnose ARDS. By embedding modalities into a common latent space rather than treating them separately, the model can draw on richer, more physiologically relevant inputs when making a prediction, improving diagnostic performance and moving towards a model that is reliable enough for decision-makers. The methodological rigor of the experiments and the heterogeneity of the training data make this an important step toward deep learning being trustworthy enough for production use in medical settings, and the joint embedding approach developed here lays the groundwork this dissertation returns to in its discussion of multimodal synthetic data generation.

2.1. Abstract

Current machine learning methods for detecting ARDS do not take full advantage of the multimodality of ARDS pathophysiology and diagnostic information. To fill that gap, we developed a multimodal deep learning model that combines a chest radiograph (x-ray), at least one hour of continuosly collected ventilator waveform (VWD) data within the first 24 hours of intubation, and EHR-derived tabular data to predict ARDS, training on data from 220 patients admitted to the ICU. To alleviate the limited dataset sizes encountered in clinical settings, the model uses pretrained encoders for the x-rays and ventilation data and trains a lightweight feature extractor on tabular data. Encoded features for a patient are combined to make a single ARDS prediction. Ablation studies for each modality assessed their effect on the model's predictive capability. The

trimodal model achieved an area under the receiver operator curve (AUROC) of 0.867 (95% CI: [0.860, 0.873]), an absolute improvement over unimodal and bimodal models, though not statistically significant. Comparator baseline methods adapted from prior published approaches achieved higher performance on several metrics when applied to our dataset, underscoring that this study's primary contribution is characterizing how classification performance changes as data modalities are incrementally combined, rather than establishing a state-of-the-art ARDS classifier. Our results demonstrate the feasibility of using deep learning to combine heterogeneous data types for a complex condition, even on a small dataset, and our framework can serve as a blueprint for future multimodal model development in similar sample size-constrained research settings.

2.2. Introduction

Acute respiratory distress syndrome (ARDS) is a severe form of hypoxemic respiratory failure that has proven challenging to diagnose and manage in a timely fashion, despite affecting up to 10% of ICU admissions and 25% of mechanically ventilated (MV) patients (Bellani et al., 2016). ARDS is associated with substantial morbidity and mortality, prolonged MV, high hospital-associated costs, and long-term physical and psychological dysfunctions (Bellani et al., 2016; M. Matthay et al., 2019). Moreover, poor outcomes in ARDS are associated with delayed and missed diagnosis and suboptimal use of evidence-based therapies, even by subspecialty-trained clinicians (Bellani et al., 2016; M. Matthay et al., 2019; Needham et al., 2015; Weiss et al., 2016), highlighting the need for tools to support early diagnosis.

The diagnostic criteria for ARDS require laboratory-based, electronic health record (EHR)-derived arterial blood gas (ABG) measurement of the degree of hypoxemia, ascertainment of mechanical ventilation settings, and the presence of bilateral opacities on chest imaging (ARDS Definition Task Force et al., 2012). This inherent multimodality of ARDS diagnostic criteria suggests that integrating multiple data sources could improve early diagnostic accuracy. Recent advances in deep learning have shown promise for early ARDS detection using single modalities, including convolutional neural networks (CNNs) on single chest radiographs (CXRs) (M. Sjoding et al., 2021) and on continuously collected ventilator waveform data (VWD) (Rehm et al., 2021). While these

unimodal approaches have shown promise, they may miss important clinical patterns that only become apparent when analyzing multiple data sources simultaneously.

Previous work has explored bimodal approaches to ARDS detection through methods that ensemble the predictions of models trained on pairs of CXR, VWD, and EHR data (Pai et al., 2022; Levy et al., 2025; Rubulotta et al., 2024). There has also been work on bimodal approaches that combine CXR and EHR data in a single deep learning model (Jabbour et al., 2022; Rubulotta et al., 2024). However, recent advances in multimodal machine learning outside of ARDS detection suggest that integrated approaches to modality fusion with more and more complex data could further improve performance. For instance, studies have demonstrated the utility of deep fusion techniques by using medical images and diagnostic notes to improve classification tasks including pneumonia detection, COVID detection, and bone abnormality detection (Y. Zhang et al., 2022). Further, studies have shown the ability to translate between cell imaging modalities by enforcing a matching objective in the latent space of autoencoders (K. D. Yang et al., 2021). These more sophisticated methods can capture both the unique contributions of each data type and the potentially important interactions between different physiologic measurements.

While bimodal approaches have shown initial promise, they have been largely limited to static types of data, such as a single measurement of hypoxemia from arterial blood gases. This relatively narrow use of data doesn't capture the dynamic nature of patients with need for mechanical ventilation leading to missed opportunities to identify the earliest patterns of disease. The complex pathophysiology of ARDS, as suggested by the diverse modalities used in the Berlin criteria, suggests that integrating VWD, CXRs, and clinical data from the EHR could capture subtle interactions between physiologic, radiographic, and clinical manifestations of the condition.

Finally, past work in ARDS detection has used thousands to tens of thousands of patients' data during training (Rubulotta et al., 2024). As collecting data from this volume of patients can be both expensive and time consuming, robust methods that can be trained with less data are needed.

Building on advances in both ARDS detection and multimodal learning, we hypothesized that combining VWD, CXRs, and EHR data into a single, trimodal deep learning model would improve

the accuracy of early ARDS detection, even when training on small, heterogeneous datasets. Specifically, we investigated whether an integrated trimodal model that uses pretrained encoders could better discriminate between ARDS and non-ARDS patients compared to unimodal and bimodal methods.

2.3. Methods

2.3.1. Cohort Description. All patient data were obtained as part of a prospective, Institutional Review Board approved study collecting data from mechanically ventilated adults admitted to the intensive care unit (ICU) at the University of California (UC) Davis Medical Center. Three critical care physicians performed retrospective chart review to identify the cause of respiratory failure in subjects from the study cohort enrolled between 2015 and 2019. Adult patients expected to require at least 24 hours of mechanical ventilation were enrolled. Patients were included if at least one hour of ventilator waveform data was collected in the first 24 hours of mechanical ventilation (non-ARDS patients) or the first 24 hours after ARDS criteria were first met (Rehm et al., 2021). The presence or absence of ARDS was determined using the Berlin consensus criteria (ARDS Definition Task Force et al., 2012) by independent clinician review, with disagreements between any two clinicians resolved by consensus chart review. For secondary analyses, subjects with ARDS were further divided into two subgroups as a function of relative ambiguity of ARDS classification: Group 1) patients with confirmed moderate or severe ARDS without chronic obstructive pulmonary disease and/or asthma (to minimize the risk of misclassifying ARDS as a result of concurrent non-ARDS acute or chronic lung disease-associated hypoxemia), chronic congestive heart failure, or severe obesity and Group 2) patients with any severity of ARDS and without exclusion of these conditions.

Causes of ARDS for patients meeting the Berlin criteria are shown in Table 2.1. Table 2.1 also includes basic demographic information and additional clinical information such as Sequential Organ Failure Assessment score, hospital length of stay, and hospital mortality.

No sample size calculation was conducted for this study, given our desire to use all patient data available in the early time period after ARDS criteria were first met; however, sample size was consistent with the range of cohort sizes in previous studies of VWD analysis (Rehm et al.,

2021) and to achieve a balanced dataset for machine learning (ML) model development. Standard ML algorithms are often biased towards the majority class, resulting in a higher misclassification rate for the minority class (Lopez et al., 2013). Because the primary purpose of this study was to assess the effects of data dimensionality on model performance and not to develop a state-of-the-art ARDS classifier, we did not selectively downsample or synthetically upsample patients to achieve a native prevalence cohort.

Characteristic	**ARDS (n = 120)**	**Non-ARDS (n = 120)**
Age (median [IQR])	57 (47-68)	62 (53-71)
Female (n [%])	53 (44)	48 (40)
Body mass index (median [IQR])	27.5 (22.4-34.7)	26.6 (22.2-33)
Sequential Organ Failure Assessment score (median 1st 24 hr [IQR])	11 (9-14.3)	8 (6-11)
Days from intubation to Berlin criteria (median [IQR])	0.11 (0.06-0.24)	—
Median PaO_2/FiO_2 first 24 hours (median [IQR])	166 (132-219.5)	311 (218-390)
Worst PaO_2/FiO_2 first 24 hours (median [IQR])	111 (73-153)	218 (135-313)
Hospital length of stay (median [IQR])	12.2 (7.3-22.7)	9 (5.3-16.9)
Hospital mortality (n [%])	57 (48)	32 (27)
Ventilator-free days in 28 days (median [IQR])	6.6 (0-21.9)	23.5 (4-26.3)
ARDS insult type (n [%])		
Pneumonia	64 (53.3)	—
Aspiration	31 (25.8)	—
Non-pulmonary sepsis	18 (15)	—
Trauma	3 (2.5)	—
Other	4 (3.3)	—

TABLE 2.1. Comparison of characteristics between ARDS and non-ARDS patients. Note: 3 non-ARDS patients were not admitted to the ICU.

2.3.2. Data Preparation. VWD representing air flow in liters per minute was collected from Puritan Bennett 840 ventilators as a one-dimenional time series at a rate of 50Hz. VWD for each patient was split into windows of 30,000 samples (10 minutes of data collection), and the number of windows for each patient was limited to the first 36 available windows (the first six hours of available data) to limit computational cost while still capturing early ARDS physiologic information. Incomplete windows were padded with 0s to create uniform window lengths. All patients had at least one hour of VWD data in the first 24 hours after Berlin criteria were first met or after the start of MV.

CXRs were obtained from a Picture Archiving and Communication System (PACS) in the form of DICOM files. For ARDS patients, the first CXR that met the Berlin criteria for ARDS was used. For non-ARDS patients, the first CXR obtained within 24 hours of intubation was used. CXRs were resized and cropped to a uniform size of 512x512 pixels.

Tabular data derived from the EHR were included from the first 24 hours after meeting Berlin criteria or the first 24 hours after the start of MV for non-ARDS patients. A full list of features used is available in Table A.1 in Appendix A. . Categorical features were one-hot encoded during data processing. Numerical features that were measured multiple times for a single patient were aggregated. For each numerical feature, we calculated the minimum, maximum, mean, and median of the feature. All numerical features were then normalized to have a mean of zero and unit variance based on the training data. Missing numerical features were filled with the mean of the feature across the current training split of the dataset. Normalization and missing data imputation of the test set were performed using the same methods, i.e., separately from the train set to prevent data leakage.

During model training, each VWD window was treated as a separate training instance, with the corresponding CXR and EHR data repeated across windows for the same patient. During inference, predictions from all windows for a single patient were averaged to produce a final classification probability.

2.3.3. Train/Test Splitting. All data splits were performed at the patient level. A 5-fold k-fold split procedure (outer folds) was used, with a further nested 5-fold split (inner folds) on each outer fold. Hyperparameter tuning was performed using a grid search on inner folds to minimize overfitting. Within each training split, data was further split 80/20 into training and validation splits. The splitting procedure can be seen in Figure 2.1.

To better represent model performance across random splits of a small dataset, the entire k-fold procedure was repeated 10 times for each modality combination, and performance was assessed using unseen data from the outer k-fold splits. This resulted in 50 total holdout sets for each model. One-hot encoding and scaling, as described in the previous section, were calculated on the training set for each split and applied to the validation and holdout sets.

2.3.4. Model Development. Each multimodal model consists of sub-classes of models: base encoders for each modality, modality-specific projection heads, and a classification head. Given a batch of patients, the multimodal model first encodes data for each patient from each modality using the base encoders. Then it passes the encoded data through modality-specific projection

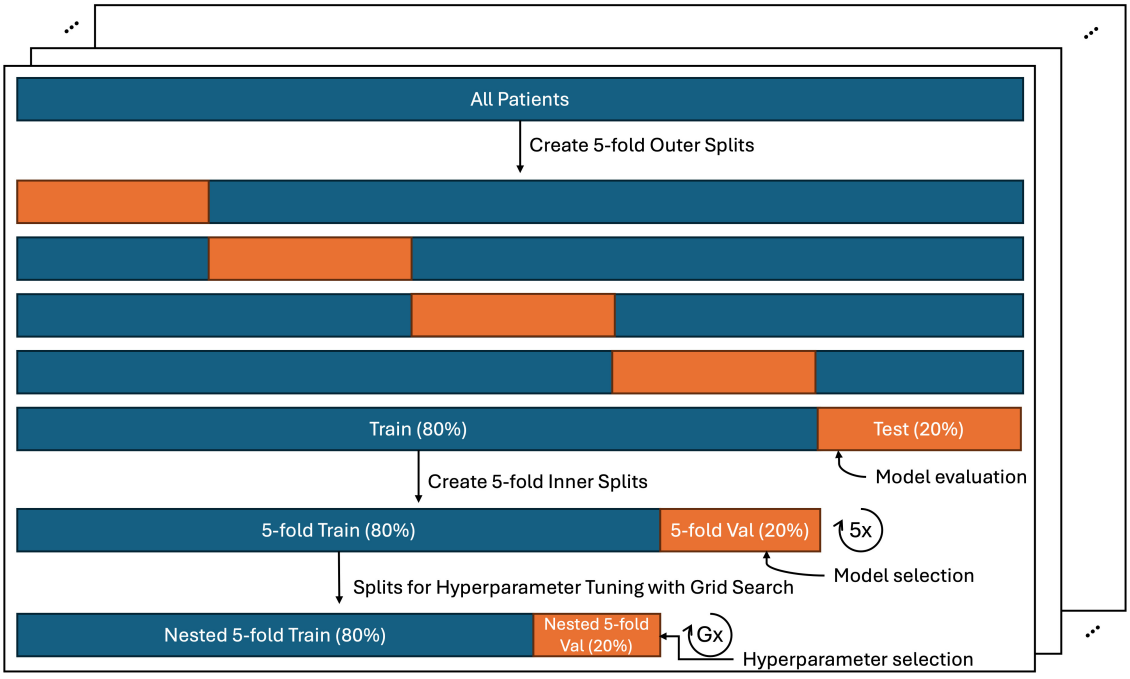


FIGURE 2.1. Splitting procedure for experimentation. Outer splits are created as a 5-fold split of the full patient dataset. Inner fold splits are created as a 5-fold split of each outer fold train split. Hyperparameter tuning was conducted by randomly splitting each inner fold train split G times, where G represents the number of hyperparameter combinations in a grid search. Final model performance was assessed on the test split of the outer folds. The full procedure was repeated 10 times.

heads. Finally, it combines the projected data via concatenation before passing it into the single classification head. Figure 2.2 shows this graphically.

Modality-specific base encoders, $Enc(\cdot)$, map an input sample, $\mathbf{x}$, to an embedding vector, $\mathbf{e} = Enc(\mathbf{x}) \in \mathbb{R}^{D_e}$, where D_e is the embedding dimension size. D_e is 128 for both CXR and VWD data, and 64 for EHR data. Base encoders are described in detail in the next section.

Projection heads, $Proj(\cdot)$, map $\mathbf{e}$ to vector $\mathbf{z} = Proj(\mathbf{e}) \in \mathbb{R}^{D_p}$. D_p is tuned as a hyperparameter for each model, with possible values of 8, 16, or 32. $Proj(\cdot)$ is a one- or two-layer perceptron network with hidden layer and output layer dimensions of D_p. $\mathbf{z}$ is normalized to the unit hypersphere in $\mathbb{R}^{D_p}$. The number of layers in $Proj(\cdot)$ is tuned as a hyperparameter.

Classification heads, $Class(\cdot)$, map $\mathbf{z}$ to two classes representing ARDS or non-ARDS. $Class(\cdot)$ is a two-layer perceptron network with input and hidden layer dimension of $D_p * m$, where m is the number of modalities, and output layer dimension of two with a softmax activation.

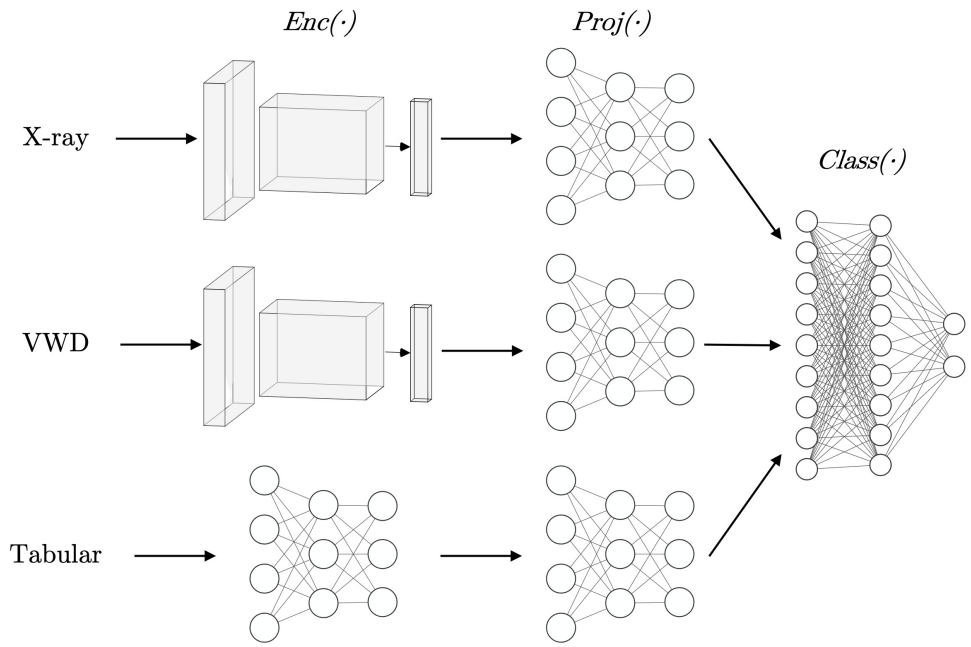


FIGURE 2.2. A schematic depiction of our trimodal architecture showing encoders, projection heads, and classification layers. X-ray and ventilator waveform encoders use a convolutional neural network architecture, while tabular encoders, projection heads, and classification networks use fully connected layers.

$Enc(\cdot)$ layers are pretrained models (CXR and VWD) or trained on each individual split (EHR). $Enc(\cdot)$ layers are frozen during downstream model training, as retraining encoder layers was found to cause instability and overtraining during early experiments. $Proj(\cdot)$ and $Class(\cdot)$ layers are jointly trained using a cross-entropy loss.

Models are trained using the Adam optimizer with an initial learning rate of 1e-4 and weight decay of 1e-6. Training is performed for a maximum of 150 epochs with early stopping based on validation loss with a patience of 3 epochs after a warm-up of 10 epochs.

We optimized the following hyperparameters during training:

- D_e: 8, 16, or 32 dimensions
- $Proj(\cdot)$ size: one- or two-layer network
- Dropout rate in $Proj(\cdot)$ and $Class(\cdot)$: 50%, 70%, or 90%

Inspired by works using a contrastive pretraining step (Y. Zhang et al., 2022; Chen et al., 2020), we also tried a version of our model where the projection layers are jointly trained to maximize the

cosine similarity of modality-specific embeddings of the same patient. However, performance on the validation set was generally better using the approach described in the paper, so this contrastive approach was not explored further.

2.3.4.1. *Base Encoders.* For each modality, we selected a pretrained encoder based on the structural similarity between the pretraining task and the characteristics of the target signal, rather than requiring the pretraining domain to match ARDS or even biomedical data directly.

The VWD-specific encoder is a CNN pretrained for audio tagging, the task of assigning one or multiple semantic labels to an audio clip (Schmid et al., 2023a). We use the pretrained model to extract intermediate features to use as our embeddings (Schmid et al., 2023b). Although VWD is a physiologic rather than acoustic signal, we selected an audio tagging encoder because both data types share an underlying structure where high-frequency, quasi-periodic one-dimensional time series data has relevant information distributed across multiple timescales. Audio tagging models are explicitly trained to extract such multiscale temporal and spectral features, making them a natural source of transferable representations for ventilator waveform data despite the difference in domain.

The CXR-specific encoder uses a pretrained model, which is a ResNet-50 backbone trained on CXRs for eight cardiopulmonary radiological tasks (Boecking et al., 2022; He et al., 2016). We remove the final projection head of the model and use the remaining network to generate CXR embeddings. We selected this encoder over a more generic ImageNet-pretrained backbone because it was pretrained specifically on chest radiographs and cardiopulmonary classification tasks, meaning its learned features are already biased toward clinically relevant pulmonary and cardiac structures rather than generic natural-image features.

The tabular-specific encoder is a model pretrained using the VIME (Value Imputation and Mask Estimation) framework (Yoon et al., 2020). This approach trains two networks on tabular data corrupted by data masks: one for data reconstruction and another for mask estimation. We use an intermediate network that serves as input to the two networks as our encoder for generating tabular embeddings. We selected VIME because it is explicitly designed for tabular data with mixed categorical and numerical features and no natural spatial or sequential structure to exploit, unlike image or waveform data. Its self-supervised mask estimation and reconstruction objectives

allow it to learn feature correlations directly from unlabeled tabular data, which is particularly useful given the comparatively small number of labeled patients available for training in this study.

These encoder choices reflect the data volume constraints imposed by our training dataset. General purpose VWD and EHR encoders aren't readily available, and training one from scratch would necessitate much more data than this study has available. Instead, we prioritized using encoders with a structural correspondence between a signal's underlying statistical properties and a pretraining task, allowing us to leverage large-scale pretrained representations despite the limited size of our ARDS-specific dataset.

2.3.4.2. *Model Baselines.* We trained two extra models that reflect comparative approaches from previous studies to ground the performance of our models. The first approach reflects a model building methodology that combines deep learning with traditional machine learning techniques. Using the training technique described by Z. Gao et al. (2024), we combine features extracted from the same CXR base encoder used in our models directly with tabular features to train a bimodal logistic regression model. The second approach reflects an alternative method for a fully end-to-end deep learning model, where the final layer of an encoder is retrained on training data. We use the training technique described by Jabbour et al. (2022) to retrain the final layers of the CXR encoder and pass tabular features through a small encoder network. Both sets of features are combined and passed through a classifier, and all networks are trained as an end-to-end bimodal model during training.

2.3.4.3. *Performance Assessment.* We investigated three individual unimodal models and four combined multimodal models. The four combined multimodal models were: VWD+CXR, VWD+tabular, CXR+tabular, and trimodal.

Final model selection for each modality combination was based on the area under the receiver operating characteristic curve (AUROC) as measured on the validation set. Model performance on unseen data from the holdout datasets was assessed by the AUROC, area under the precision-recall curve (AUPRC) and typical confusion matrix statistics. Because our holdout set's class balance (50/50) differs from the class balance that would be expected in clinical settings (75/25), all AUPRC are corrected values using the fact that true positive rates (TPRs) and false positive rates (FPRs) are invariant to class prevalence, allowing precision to be recalculated at any target

prevalence from the observed ROC curve (Saito et al., 2015). Precision at each threshold, t, was recomputed as

$$\text{Precision}'(t) = \frac{\text{TPR}(t)\pi'}{\text{TPR}(t)\pi' + \text{FPR}(t)(1-\pi')}$$

for target prevalence $\pi' = 0.25$.

Because the 5-fold cross-validation was repeated across 10 random initializations, the resulting performance estimates are not independent: overlapping training and test partitions across repeats induce correlation that violates the independence assumption of standard paired significance tests, leading to underestimated variance and inflated Type I error rates (Dietterich, 1998). To account for this, we used the corrected resampled paired t-test of Nadeau et al. (2003), which adjusts the variance estimate for the correlation introduced by repeated k-fold resampling. This test was applied pairwise to all model comparisons, and the resulting p-values were adjusted for multiple comparisons using the Holm-Bonferroni procedure Holm, 1979, which controls the family-wise error rate while being less conservative than the standard Bonferroni correction.

Performance metrics are reported with 95% confidence intervals calculated with python's SciPy library (Virtanen et al., 2020), where we bootstrap the CI by resampling repeats (not raw folds) with replacement 5,000 times to respect the within-repeat correlation. Full raw and adjusted p-values are in Table 2.2. Model calibration was assessed quantitatively using Brier score and descriptively using calibration curves plotting the predicted probability of ARDS against the observed prevalence of ARDS across predicted probability strata.

All models were trained on a single NVIDIA Tesla T4 GPU with 16 GB GDDR6 memory.

2.4. Results

2.4.1. Overall Model Performance. Across all evaluated modality combinations, the trimodal model achieved the highest overall performance across confusion matrix measures with an AUROC of 0.867 (CI: [0.860, 0.873]) and an overall accuracy of 78.2% (CI: [77.2%, 79.0%]) when averaged across all holdout datasets. The trimodal model did not achieve a statistically significant improvement over either unimodal models or bimodal models(adjusted p>0.05). Table 2.3 shows the full complement of discrimination metrics averaged across all holdout datasets for models trained on each modality combination. When compared against previously published multimodal

Modality 1	Modality 2	Raw P-value	Adjusted P-value
Tabular	Z. Gao et al. (2024)	0.131303	1
Tabular	Jabbour et al. (2022)	0.178002	1
Tabular	Trimodal	0.466441	1
Tabular	VWD	0.028783	0.834694
Tabular	VWD+Tabular	0.959449	1
Tabular	VWD+X-ray	0.641703	1
Tabular	X-ray	0.089234	1
Tabular	X-ray+Tabular	0.697264	1
Z. Gao et al. (2024)	Jabbour et al. (2022)	0.785909	1
Z. Gao et al. (2024)	Trimodal	0.63108	1
Z. Gao et al. (2024)	VWD	0.000337	0.012142
Z. Gao et al. (2024)	VWD+Tabular	0.236398	1
Z. Gao et al. (2024)	VWD+X-ray	0.057884	1
Z. Gao et al. (2024)	X-ray	0.002297	0.078102
Z. Gao et al. (2024)	X-ray+Tabular	0.294212	1
Jabbour et al. (2022)	Trimodal	0.765963	1
Jabbour et al. (2022)	VWD	0.001313	0.045968
Jabbour et al. (2022)	VWD+Tabular	0.285648	1
Jabbour et al. (2022)	VWD+X-ray	0.088652	1
Jabbour et al. (2022)	X-ray	0.005477	0.175263
Jabbour et al. (2022)	X-ray+Tabular	0.343085	1
Trimodal	VWD	0.004919	0.162333
Trimodal	VWD+Tabular	0.445398	1
Trimodal	VWD+X-ray	0.201091	1
Trimodal	X-ray	0.008929	0.276792
Trimodal	X-ray+Tabular	0.711163	1
VWD	VWD+Tabular	0.031161	0.872519
VWD	VWD+X-ray	0.100526	1
VWD	X-ray	0.668857	1
VWD	X-ray+Tabular	0.014631	0.43892
VWD+Tabular	VWD+X-ray	0.668144	1
VWD+Tabular	X-ray	0.106485	1
VWD+Tabular	X-ray+Tabular	0.771446	1
VWD+X-ray	X-ray	0.250164	1
VWD+X-ray	X-ray+Tabular	0.468214	1
X-ray	X-ray+Tabular	0.065533	1

TABLE 2.2. The statistical significance between average AUROC, as detailed in Section 2.3.4.3, between models trained on different modalities.

approaches (Z. Gao et al., 2024; Jabbour et al., 2022), however, these comparator methods achieved higher performance on several metrics, which we address in the Discussion.

2.4.2. Modality Contribution Analysis. We observed consistent performance gains in unimodal models across almost all measures of discrimination when adding either a second or third modality, with the trimodal model showing the highest overall performance (Table 2.3). The model combining VWD+CXR data yielded substantially better performance than either modality alone (VWD+CXR model AUROC of 0.816 (CI: [0.808, 0.827]), compared to unimodal VWD and CXR models with AUROCs of 0.743 (CI: [0.740, 0.747]) and 0.762 (CI: [0.753, 0.770]), respectively. While the gains were more modest, adding either VWD or CXR to the Tabular model also resulted in performance gains across all metrics except specificity.

2.4.3. Performance Stratification by Patient Group. Model performance varied notably between Group 1 (more clear-cut ARDS cases) and Group 2 (less clear-cut cases). In Group 1, the trimodal model achieved the highest AUROCs of our models for both Group 1 (0.95, CI: [0.94, 0.96]) and Group 2 0.79 (CI: [0.78, 0.80]) ARDS cases. We observed a consistent performance gap between Group 1 and Group 2 ARDS cases across all modalities, but the gap was not uniform (Table 2.4). For unimodal models, X-ray data showed the largest performance change (ΔAUROC = 0.18), followed by VWD data (ΔAUROC = 0.17), and then tabular data (ΔAUROC = 0.17). In bimodal models, VWD+Tabular had the smallest performance decrease (ΔAUROC = 0.14) followed by CXR+Tabular (ΔAUROC = 0.17) had VWD+ CXR, which had the largest drop for bimodal models (ΔAUROC = 0.19).

Modality	Cohort	Accuracy	AUROC	Precision	Sensitivity	Specificity	AUPRC
Tabular	All	0.766 ± [0.759, 0.773]	0.834 ± [0.829, 0.839]	0.784 ± [0.775, 0.796]	0.740 ± [0.722, 0.758]	0.793 ± [0.782, 0.808]	0.617 ± [0.597, 0.635]
X-ray	All	0.688 ± [0.678, 0.698]	0.762 ± [0.753, 0.770]	0.682 ± [0.647, 0.711]	0.693 ± [0.662, 0.719]	0.683 ± [0.649, 0.717]	0.584 ± [0.561, 0.605]
VWD	All	0.685 ± [0.681, 0.690]	0.743 ± [0.740, 0.747]	0.690 ± [0.685, 0.696]	0.705 ± [0.698, 0.712]	0.665 ± [0.656, 0.676]	0.504 ± [0.492, 0.516]
VWD+X-ray	All	0.735 ± [0.726, 0.743]	0.816 ± [0.808, 0.827]	0.751 ± [0.741, 0.762]	0.729 ± [0.710, 0.748]	0.740 ± [0.724, 0.755]	0.642 ± [0.623, 0.661]
VWD+Tabular	All	0.756 ± [0.741, 0.772]	0.836 ± [0.824, 0.848]	0.772 ± [0.756, 0.786]	0.760 ± [0.741, 0.783]	0.755 ± [0.733, 0.775]	0.646 ± [0.620, 0.671]
X-ray+Tabular	All	0.764 ± [0.757, 0.772]	0.849 ± [0.842, 0.855]	0.781 ± [0.766, 0.795]	0.753 ± [0.736, 0.769]	0.776 ± [0.754, 0.798]	0.711 ± [0.694, 0.727]
Trimodal	All	0.782 ± [0.772, 0.790]	0.867 ± [0.860, 0.873]	0.797 ± [0.785, 0.808]	0.776 ± [0.765, 0.787]	0.789 ± [0.774, 0.803]	0.740 ± [0.725, 0.755]
Z. Gao et al. (2024)	All	0.785 ± [0.781, 0.789]	0.888 ± [0.883, 0.893]	0.793 ± [0.786, 0.800]	0.782 ± [0.775, 0.790]	0.788 ± [0.778, 0.798]	0.781 ± [0.769, 0.793]
Jabbour et al. (2022)	All	0.804 ± [0.794, 0.813]	0.879 ± [0.874, 0.884]	0.805 ± [0.789, 0.818]	0.813 ± [0.806, 0.820]	0.795 ± [0.773, 0.812]	0.743 ± [0.730, 0.757]

TABLE 2.3. Average metrics and lower and upper 95% confidence intervals (in square brackets) for models trained on each modality combination. Additionally, results are shown for models trained using the methods described by Z. Gao et al. (2024) and Jabbour et al. (2022).

Modality	Cohort	Accuracy	AUROC	Precision	Sensitivity	Specificity	AUPRC
Tabular	Group 1	0.85 ± [0.84, 0.86]	0.93 ± [0.92, 0.94]	0.86 ± [0.84, 0.88]	0.82 ± [0.80, 0.84]	0.88 ± [0.86, 0.90]	0.85 ± [0.82, 0.87]
X-ray		0.79 ± [0.78, 0.80]	0.86 ± [0.84, 0.88]	0.80 ± [0.77, 0.83]	0.72 ± [0.70, 0.74]	0.84 ± [0.81, 0.86]	0.79 ± [0.77, 0.81]
VWD		0.77 ± [0.77, 0.78]	0.84 ± [0.84, 0.84]	0.78 ± [0.77, 0.79]	0.78 ± [0.77, 0.79]	0.77 ± [0.76, 0.78]	0.69 ± [0.68, 0.70]
VWD+X-ray		0.84 ± [0.83, 0.85]	0.92 ± [0.91, 0.93]	0.89 ± [0.87, 0.91]	0.79 ± [0.76, 0.81]	0.89 ± [0.87, 0.91]	0.86 ± [0.85, 0.87]
VWD+Tabular		0.82 ± [0.80, 0.84]	0.91 ± [0.90, 0.92]	0.84 ± [0.82, 0.85]	0.83 ± [0.80, 0.86]	0.82 ± [0.80, 0.84]	0.81 ± [0.79, 0.83]
X-ray+Tabular		0.86 ± [0.85, 0.87]	0.94 ± [0.93, 0.95]	0.88 ± [0.86, 0.90]	0.83 ± [0.81, 0.84]	0.89 ± [0.87, 0.91]	0.89 ± [0.88, 0.90]
Trimodal		0.87 ± [0.86, 0.88]	0.95 ± [0.94, 0.96]	0.91 ± [0.89, 0.93]	0.83 ± [0.82, 0.85]	0.92 ± [0.90, 0.94]	0.91 ± [0.90, 0.92]
Z. Gao et al. (2024)		0.88 ± [0.87, 0.89]	0.97 ± [0.97, 0.97]	0.88 ± [0.87, 0.89]	0.89 ± [0.88, 0.91]	0.88 ± [0.87, 0.89]	0.92 ± [0.91, 0.93]
Jabbour et al. (2022)		0.91 ± [0.90, 0.92]	0.97 ± [0.96, 0.98]	0.93 ± [0.91, 0.95]	0.88 ± [0.86, 0.90]	0.94 ± [0.92, 0.96]	0.95 ± [0.94, 0.96]
Tabular	Group 2	0.70 ± [0.69, 0.71]	0.76 ± [0.75, 0.77]	0.73 ± [0.71, 0.75]	0.67 ± [0.65, 0.69]	0.73 ± [0.71, 0.75]	0.52 ± [0.49, 0.55]
X-ray		0.62 ± [0.60, 0.63]	0.68 ± [0.67, 0.69]	0.61 ± [0.57, 0.64]	0.67 ± [0.63, 0.71]	0.56 ± [0.52, 0.60]	0.52 ± [0.49, 0.55]
VWD		0.62 ± [0.61, 0.63]	0.67 ± [0.67, 0.67]	0.62 ± [0.61, 0.63]	0.65 ± [0.64, 0.66]	0.59 ± [0.58, 0.60]	0.43 ± [0.42, 0.44]
VWD+X-ray		0.66 ± [0.65, 0.67]	0.73 ± [0.71, 0.75]	0.66 ± [0.64, 0.68]	0.68 ± [0.65, 0.71]	0.64 ± [0.62, 0.66]	0.52 ± [0.50, 0.55]
VWD+Tabular		0.71 ± [0.69, 0.73]	0.77 ± [0.76, 0.78]	0.72 ± [0.70, 0.74]	0.71 ± [0.69, 0.74]	0.71 ± [0.68, 0.74]	0.56 ± [0.53, 0.59]
X-ray+Tabular		0.69 ± [0.68, 0.70]	0.77 ± [0.76, 0.78]	0.72 ± [0.70, 0.73]	0.70 ± [0.67, 0.73]	0.69 ± [0.66, 0.72]	0.61 ± [0.58, 0.64]
Trimodel		0.72 ± [0.71, 0.73]	0.79 ± [0.78, 0.80]	0.72 ± [0.70, 0.73]	0.74 ± [0.72, 0.76]	0.71 ± [0.68, 0.74]	0.63 ± [0.61, 0.65]
Z. Gao et al. (2024)		0.72 ± [0.71, 0.73]	0.82 ± [0.81, 0.83]	0.74 ± [0.73, 0.75]	0.72 ± [0.71, 0.74]	0.72 ± [0.70, 0.74]	0.70 ± [0.68, 0.72]
Jabbour et al. (2022)		0.72 ± [0.70, 0.74]	0.79 ± [0.78, 0.80]	0.72 ± [0.70, 0.74]	0.76 ± [0.74, 0.78]	0.68 ± [0.65, 0.71]	0.65 ± [0.64, 0.66]

TABLE 2.4. Average metrics and lower and upper 95% confidence intervals (in square brackets) for models trained on each modality combination, with results broken down by patient group. Additionally, results are shown for models trained using the methods described by Z. Gao et al. (2024) and Jabbour et al. (2022).

2.4.4. Model Calibration. Qualitatively, the trimodal model was generally well-calibrated across its range of predicted probabilities in the holdout datasets, with a small overprediction of ARDS risk toward the lower range of predictions and similarly small underprediction of risk in the upper range (Figure 2.3, with all modality combinations in Figure A.1 of Appendix A). Quantitatively, the trimodal model achieved a Brier score of 0.17 (Brier scores for all other models can be found in Table A.2 of Appendix A). Table 2.5 shows the performance of the trimodal model as assessed by confusion matrix statistics across deciles of model classification thresholds. Note that the number of patients in each decile of predicted probability was not uniform and models made no predictions with probability <0.2 or >0.8 (grey histogram bars, Figure 2.3).

2.5. Discussion

In this study, we examined the feasibility of multimodal deep learning for ARDS detection on heterogeneous datasets by training small models on top of larger pretrained models with different combinations of modalities. We found that small models built on top of the generalized embeddings of larger pretrained models can reliably classify ARDS in ICU patients using data collected in a

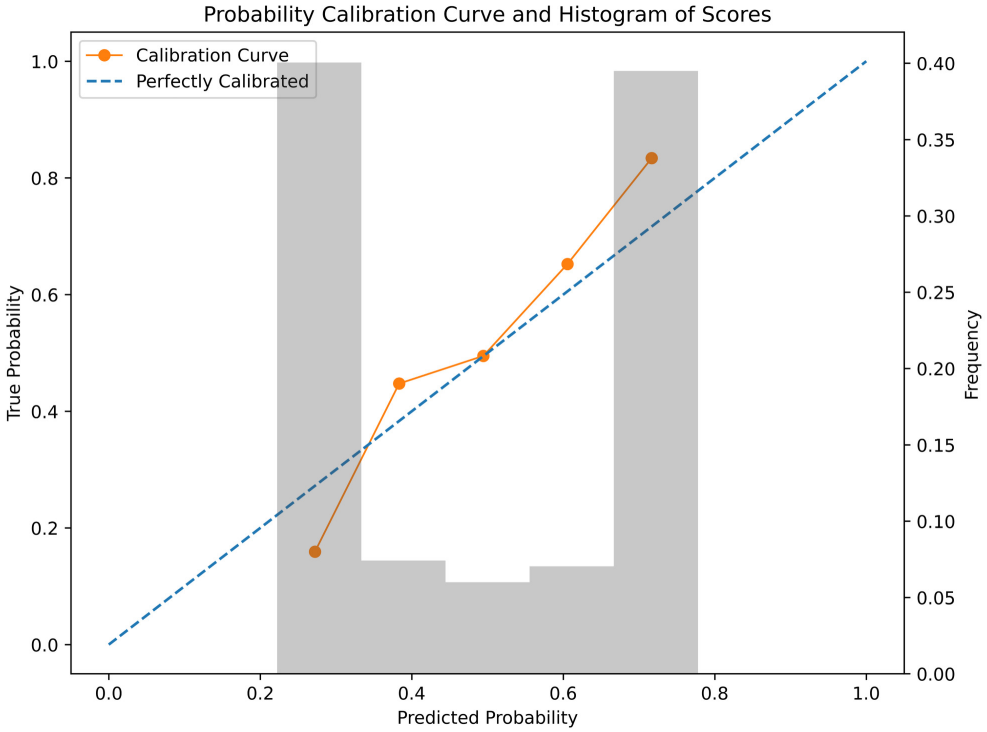


FIGURE 2.3. A calibration curve (in orange) for our trimodal model showing predicted probability on the x-axis and true probability on the left y-axis. Overlayed with the calibration curve is the prediction distribution (in gray). The distribution shows the prediction frequency, on the right y-axis, at each probability decile, the x-axis.

retrospective study. We also found that models trained on multiple modalities tend to outperform models trained on subsets of those modalities, though not statistically significantly. Specifically, we showed that trimodal models tended to outperform bimodal models and that bimodal models tended to outperform unimodal models. We found all models to be generally well calibrated across the range of predicted probabilities.

Our findings build upon recent research (Rubulotta et al., 2024; Levy et al., 2025) that collectively supports the potential utility of using machine learning to predict the presence of ARDS. While not outperforming comparable training setups by Z. Gao et al. (2024) and Jabbour et al. (2022), our work extends previous work on machine learning-based ARDS detection in two ways. First, we tested the combinatorial addition of three clinical data sources, which, to our knowledge,

Threshold	Accuracy	Precision	Sensitivity	Specificity
0.0	—	—	—	—
0.1	—	—	—	—
0.2	—	—	—	—
0.3	0.75	0.71	0.89	0.61
0.4	0.78	0.76	0.83	0.72
0.5	0.78	0.78	0.78	0.78
0.6	0.77	0.82	0.71	0.83
0.7	0.74	0.86	0.59	0.89
0.8	—	—	—	—
0.9	—	—	—	—
1.0	—	—	—	—

TABLE 2.5. Average trimodal model performance characteristics across predicted probability thresholds.

is the first trimodal deep learning approach to ARDS classification. Second, while some prior approaches collect EHR data over windows exceeding 48 hours (Z. Gao et al., 2024), our framework constrains all modality data collection to a 24-hour window following intubation or the development of ARDS, demonstrating the feasibility of training a model on a narrower, more timely data window.

Although this study was methodological in intent, a few of the patterns we observed may be informative if future work were to pursue clinical translation. We found that adding either VWD or CXRs to EHR data achieved comparable performance gains. Because chest imaging is ordered relatively infrequently in the general mechanically ventilated population, a combined EHR+VWD approach could be better suited to a model intended for broad, frequent screening across all MV patients (Zeiberg et al., 2019; Levy et al., 2025), pointing to a possible role for continuous physiologic monitoring signals in multimodal models of this kind (Zheng et al., 2021; Hannun et al., 2019; Gustafsson et al., 2022; Tjepkema-Cloostermans et al., 2019). By contrast, CXR availability is less likely to be a practical constraint specifically within the population of patients who go on to develop ARDS, since they typically receive at least one chest radiograph within the first 24 hours of mechanical ventilation as part of standard care. This suggests CXR-inclusive models could

still have a role in more targeted, rather than universal, screening applications, for example those triggered by the availability of each newly performed CXR.

Our finding of additional, albeit modest, gains in performance with the use of three data sources suggests that the use of distinct but complementary data types may enable more effective classifiers to be developed. While our work demonstrates the potential utility of adding a third, more complex modality to an ARDS detection model and provides the flexibility to add more, our results highlight the need for additional research into other difficult-to-diagnose conditions and those with multimodal diagnostic criteria. This is especially true considering our results indicating that existing bimodal approaches from Z. Gao et al. (2024) and Jabbour et al. (2022) still outperform the best-performing trimodal model trained using the approach explored in this study. Additional research could also help determine whether the added complexity and computational costs of multimodal model development are consistently beneficial, and whether certain conditions benefit more or less from the inclusion of certain data types.

We were surprised by our analysis of model performance in subgroups of "pure" ARDS (Group 1) versus patients with ARDS and concurrent chronic lung disease, congestive heart failure, or severe obesity (Group 2). We hypothesized that we would see a relatively larger improvement in measures of discrimination with the bimodal and trimodal models in Group 2 due to the greater medical complexity of that subgroup, where adding additional information content might help models to learn ARDS from non-ARDS cases. While bimodal and trimodal models saw consistent, and in many cases substantial, improvements in performance in Group 1 patients, Group 2 patients did not see consistent gains with additional data types. Our findings, if replicated, suggest that multimodal classification of conditions like ARDS may have gaps. This could be in part due to the known imperfect validity of ARDS diagnostic criteria, which have undergone multiple revisions over decades (Ranieri et al., 2012; Thille et al., 2013; Ferguson et al., 2004; M. A. Matthay et al., 2024). Thus, the substantial physiologic overlap between clear and ambiguous cases (M. W. Sjoding et al., 2019) may not simply be addressed with a "more is better" strategy and instead may require subgroup-specific approaches to model development, including but not limited to feature importance analyses to identify which data types contribute most to performance in Group 2. Validation in

multiple clinical centers would shed light on the utility of the "more is better" approach and would help determine if our results are an artifact of the three modalities chosen and/or local practices.

Our work also extends prior research in how it used pretrained models. Our approach used two pretrained models, one for CXR data and one for VWD, and then trained smaller models on top of the pretrained models, which are frozen during training. In the past, similar ARDS models have employed training strategies like pretraining a model on a general CXR dataset and finetuning it on a smaller ARDS dataset to make unimodal predictions (M. Sjoding et al., 2021), or using pretrained models to extract features that are then combined with tabular features to train traditional ML models (N. Reamaroon et al., 2021; Afshin-Pour et al., 2023; Gandomi et al., 2024; Levy et al., 2025; Z. Gao et al., 2024). Purely deep learning-based works with a training strategy most similar to ours train only the final layers of a pretrained model on ARDS-specific data (Pai et al., 2022; Jabbour et al., 2022). However, Pai et al. (2022)built their multimodal model by ensembling the predictions of unimodal models, while our modalities are combined into a single model. Moreover, retraining layers of large models comes with speed drawbacks. The average model training time for our approach was about 20s while the average training time for the approach outlined by Jabbour et al. (2022) was 200s in our testing. Finally, we could not find any pre-existing ARDS detection work that uses pretrained models for waveform data.

Our use of pretrained models also allowed us to extend prior research in the number of patients used during training. Z. Gao et al. (2024) represent a particularly strong comparable work, employing a rigorous multimodal framework that achieves impressive performance (AUROC 0.856 in their original study) in only about 2s of training time during our testing and most closely mirrors our own approach among existing studies. When applied to our dataset, their method achieved an AUROC of 0.888, outperforming our trimodal model's observed performance (AUROC 0.867), which was instead similar to other published studies (Thille et al., 2013; Xiong et al., 2025). This gap suggests that architectures explicitly optimized for predictive performance may still hold an advantage over our framework, even though our approach was designed to prioritize methodological transparency and feasibility under small-sample, multimodal conditions rather than to maximize discrimination. Further, the level of performance reached by both approaches on such a small sample space suggests a potential for improved performance in larger datasets if they are available.

2.5.1. Technical Innovation and Broader Impact. Our approach serves as a proof of concept for adapting ideas from the broader world of ML and multimodal ML research. Using a small classification network on top of a larger, frozen pretrained network has become common practice for image classification (Chen et al., 2020). The concept of using a projection network to reduce the dimensionality of a feature space is common for networks like autoencoders (Schmidhuber, 2015), and bottlenecks have also been used for multimodal networks outside of the biomedical space (Nagrani et al., 2021). By combining many of these ideas, we showed the viability of multimodal deep learning on small heterogeneous datasets. Deep learning is notoriously data-hungry, and our application of pretrained models, combined with carefully selected architectures, demonstrated the feasibility of developing accurate, well-calibrated models, even with limited training data. During model development, we also explored more sophisticated fusion strategies, including contrastive learning approaches (K. D. Yang et al., 2021; Y. Zhang et al., 2022), but found that these methods did not improve performance over late fusion in our dataset, possibly due to overfitting on the small sample size. The fact that a simpler fusion strategy achieved strong performance may itself be informative for the community, particularly for groups seeking to deploy multimodal models in resource-constrained clinical settings where model complexity can be a practical barrier.

Our approach to handling multimodal data could serve as a template for other conditions where diagnosis relies on the integration of multiple data types, particularly in critical care settings where clinicians already routinely synthesize data from multiple sources. The architecture's modality-specific encoders and joint classification head could, in principle, be extended to incorporate additional modalities, such as other imaging types, other one-dimensional physiologic signals, or clinical notes. More specifically, because the framework is designed to perform well on small multimodal datasets, it may be a useful starting point for future work in clinical domains where data scarcity is often a barrier to applying advanced ML techniques.

2.5.2. Limitations. This work was designed as a methodological investigation into the relationship between the number of data modalities used in training and classifier performance on a small, heterogeneous ARDS dataset, rather than as a validation of a clinical diagnostic or screening tool. Accordingly, our results should be interpreted as evidence for the feasibility of multimodal

architectures in data-constrained biomedical settings, not as evidence of readiness for clinical deployment.

Our study has several important limitations. First, the single-center nature of our study limits our ability to assess generalizability across different clinical settings and patient populations. ARDS, by definition, occurs across a wide spectrum of inciting conditions, and thus, our ARDS patient population may not be generalizable to all ARDS cases. Different institutions may have different ventilator management strategies, although our institution follows ARDSnet protocols. ARDS patients may also have been imaged at different intervals. These factors could affect the relative contributions of each data type to multimodal model classification performance. Second, while our sample size was sufficient to demonstrate the potential of our approach, a larger dataset would be needed to fully validate the model's performance and ensure robust learning of complex cross-modal patterns. This is particularly important for the more complex ARDS cases in Group 2, where more examples would help better understand the relative contributions of each data type to model performance in less clear-cut cases and in key ARDS subgroups (Calfee et al., 2014; Valda Toro et al., 2024).

Third, our retrospective study design was chosen to allow investigation of the effects of multimodal deep learning on ARDS classification but doesn't address several challenges that would arise in real-world implementation. These include the need for real-time data processing and integration with clinical workflows. Our design also does not address the challenge of handling missing or delayed data that a model designed for clinical application would need to account for, nor does it analyze or take into account the effect that the relative time of multimodal data collection could have on a model's predictions.

While our model focuses on data available within the first 24 of meeting diagnostic criteria, ARDS remains challenging to diagnose even after this time(Bellani et al., 2016). Models focused on classification after this 24-hour window, therefore, may still have utility. However, early recognition and frequent reassessment would facilitate the use of evidence-based therapies in conditions like ARDS where earlier treatment is thought to improve survival (Needham et al., 2015).

Finally, while our models showed good calibration within their ranges of predicted values, the absence of predictions in certain probability ranges (0.0-0.2 and 0.8-1.0) suggests some limitations

in the models' ability to express very high or very low confidence. This could be addressed in future work through improved model architecture or training approaches, including the potential inclusion of clinician uncertainty regarding the target label into the training process (Narathip Reamaroon et al., 2019). Future work aimed at clinical translation would need to address these considerations directly.

2.6. Conclusion

Our results demonstrate the potential of multimodal deep learning to progressively improve ARDS classification by increasing the number of unique data types. Moreover, our methodological approach was able to produce performant models using small training datasets. However, we found heterogeneous performance with different combinations of modalities and in subgroups of ARDS patients, highlighting that more may not always be better. By showing that the integration of multiple data modalities can improve model performance, even in ambiguous cases, our work contributes to the broader goal of developing AI systems that can meaningfully support clinical decision-making in complex conditions and high-stakes clinical environments.

2.7. Acknowledgments

The authors acknowledge support from NIH grant R01HL16351. The authors also acknowledge Anna Liu and Irene Cortes-Puch, MD for their work on data preparation, and the UC Davis Health Information Technology Data Provisioning Core and Health Computing Core.

2.8. Ethics Statement

This research involving human subjects was conducted in compliance with relevant laws and institutional guidelines. The study protocol was reviewed and approved by the UC Davis Institutional Review Board (IRB) under reference numbers 647002-8 and 19-28933.

All procedures involving the collection and analysis of patient data, including chest radiographs, ventilator waveform data, and electronic health record information, were performed in accordance with institutional policies for human subjects research and data privacy protection.

Informed consent was obtained from all patients or their legally authorized representatives prior to data collection and analysis. Participants were informed of the research objectives, data usage, and their rights regarding participation.

All patient data were de-identified in accordance with HIPAA Privacy Rule requirements to protect patient privacy and confidentiality. No personally identifiable information was used in the analysis or will be reported in any publications resulting from this research.

The research team completed required training in human subjects protection and maintained strict protocols for data security and confidentiality throughout the study period.

CHAPTER 3

Optimal Transport Learning through Synchronized Forward-Backward Interpolants

This chapter is based on as-yet unpublished work that is a collaboration with Shizhou Xu, Jiachen Liu, and Thomas Strohmer from the UC Davis Department of Mathematics (at the time of the work) and Shih-Hsin Wang, Yuhao Huang, and Bao Wang from the University of Utah (at the time of the work). Building on the joint embedding approach developed in Chapter 2, this chapter turns from diagnosis to generation, and specifically to the problem of synthesizing chest x-rays that plausibly reflect the presence or absence of a given condition. The chapter frames synthetic x-ray generation as a transport problem: isolating what in an image distinguishes a condition and learning to add or remove it while leaving the rest of the image unchanged. While existing optimal transport methods for this kind of unpaired generation typically rely on expensive solvers, minibatch approximations, or restrictive parameterizations, this work develops a bidirectional dynamic optimal transport framework that learns transport trajectories directly, without paired samples or precomputed matching, while providing theoretical guarantees on how closely the learned map approximates the true optimal map. That theoretical grounding is central to the chapter's contribution to this dissertation's broader aim in that a synthetic image is only useful to medical decision-makers if there is a principled reason to trust it. This chapter's rigor around how the model learns to generate is what makes that trust possible. Together with the joint embedding work of Chapter 2, this chapter lays the second piece of groundwork necessary for coordinated, multimodal synthetic data generation.

3.1. Abstract

Learning high-dimensional optimal transport (OT) maps from unpaired source and target samples is challenging, since existing methods often rely on expensive OT solvers, minibatch coupling

approximations, iterative trajectory straightening, or restrictive convex-potential parameterizations. We propose a bidirectional dynamic OT framework that learns transport trajectories without paired samples or precomputed OT matching. The key idea is to construct synchronized forward and backward McCann-type interpolants that minimize both the expected trajectory length and a consistency loss, which encourages invertibility and intermediate marginal distribution matching, while enforcing marginal agreement on both source and target distributions through adversarial discrepancy. Theoretically, we derive a consistency result showing that the interaction between the transport L^2 loss and the proposed consistency loss provides a practically tractable theoretical guarantee on the closeness between the learned OT map and the optimal map under mild assumptions. Empirically, we demonstrate accurate and stable transport learning on MNIST, CelebA, single-cell perturbation data, and chest x-ray images.

3.2. Introduction

Learning meaningful transformations between unpaired high-dimensional distributions is a fundamental challenge in modern machine learning. Such problems arise in image-to-image translation (Zhu et al., 2017), domain adaptation (Courty et al., 2017), and single-cell perturbation modeling (Bunne et al., 2023), where ground-truth correspondences are inherently unobserved and matching endpoint marginals alone is insufficient. The goal is instead to recover a structurally meaningful map that reflects the underlying geometry or dynamics. Optimal Transport (OT) provides a principled framework for this task through a minimal-effort prior (Villani, 2009; Santambrogio, 2015; Peyré et al., 2019): under quadratic cost, the Brenier map is the gradient of a convex potential (Brenier, 1991), and the induced McCann displacement interpolation defines a Wasserstein geodesic between the two distributions (McCann, 1997). Equivalently, the Benamou–Brenier dynamic formulation characterizes this path as the minimum-kinetic-energy solution of a continuity equation (Benamou et al., 2000).

Despite this appealing characterization, learning OT maps remains difficult in high dimensions. Classical computational OT methods, including entropic Sinkhorn iterations (Cuturi, 2013; Peyré et al., 2019), are primarily discrete, can be expensive at large sample sizes, and do not directly provide parametric maps for out-of-sample generalization. Recent flow-matching methods improve

scalability by using prescribed couplings between source and target samples (Lipman et al., 2022). In particular, minibatch OT couplings are used to straighten flow trajectories and improve few-step generation (Pooladian et al., 2023; Tong et al., 2023). However, these couplings are batch-dependent approximations of the population OT plan and may introduce both computational overhead and statistical bias (Fatras et al., 2019).

Neural OT methods instead seek to learn transport maps or potentials directly from samples. Dual and minimax formulations, including Wasserstein adversarial objectives and neural OT methods (Arjovsky et al., 2017; Korotin et al., 2023), are principled but can be sensitive to optimization and provide limited direct supervision of the intermediate transport path. Brenier-inspired approaches represent the quadratic OT map as the gradient of a convex potential, often parameterized by Input Convex Neural Networks (Brenier, 1991; Amos et al., 2017). While this preserves the convex-potential structure required by Brenier's theorem, ICNN parameterizations can be restrictive and difficult to optimize compared with standard neural architectures. Optimal Flow Matching (Kornilov et al., 2024) is closely related in its goal of learning straight OT trajectories, but still belongs to the broader class of methods whose practical performance depends on surrogate coupling or potential-learning mechanisms. As a result, existing methods often trade off scalability, architectural flexibility, optimization stability, and fidelity to the population OT solution.

Very recent work further highlights the importance of coupling design in flow-based generative modeling. Semi-discrete OT coupling methods replace minibatch OT with dataset-level semi-discrete formulations to improve scalability and trajectory straightness (Mousavi-Hosseini et al., 2025; Kong et al., 2026), while hierarchical or dependent-coupling approaches modify the coupling structure to improve few-step or multistage generation (Yichi Zhang et al., 2025; Ma et al., 2026). These methods reinforce the importance of coupling structure, but they still construct couplings or interpolants through external pairing mechanisms. In contrast, our method obtains intermediate supervision intrinsically by synchronizing forward and backward map-induced interpolants, without precomputing pairings or solving minibatch/semi-discrete OT subproblems during training.

3.2.1. Our Contributions. We propose a new framework for learning the quadratic OT map directly from unpaired samples, without iterative Reflow-style retraining (Xingchao Liu et al., 2023),

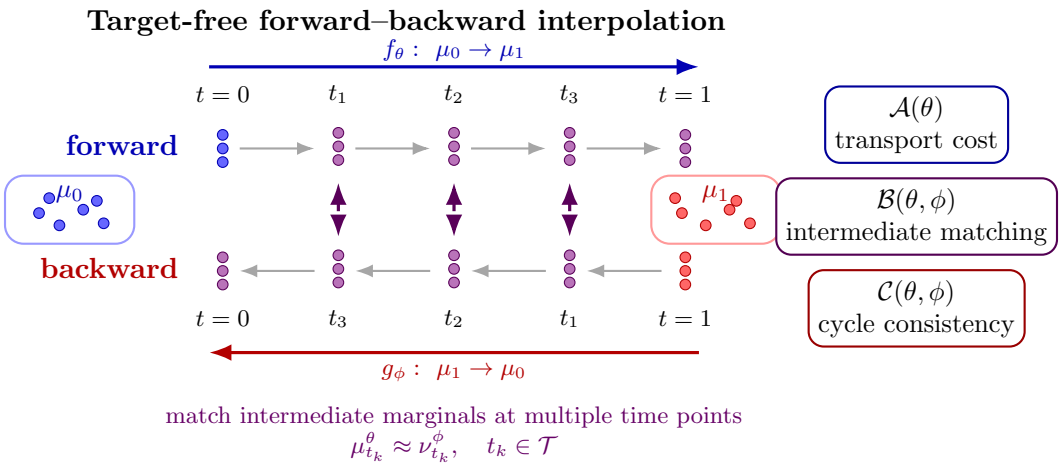


FIGURE 3.1. **Target-free OT learning via synchronized forward–backward interpolants.** The forward map f_θ transports μ_0 from left to right, while the backward map g_ϕ transports μ_1 from right to left. At multiple intermediate times, the corresponding forward and backward induced marginals are matched, providing target-free supervision of the transport path.

minibatch or semi-discrete OT coupling solvers (Pooladian et al., 2023; Tong et al., 2023; Mousavi-Hosseini et al., 2025; Kong et al., 2026), or restrictive convex-potential parameterizations (Amos et al., 2017). Our method learns two maps simultaneously: a forward transport map from source to target and a backward map from target to source. From these maps, we construct synchronized forward and backward interpolants that evolve from opposite endpoints and are required to agree at intermediate times. This creates intrinsic, target-free supervision along the transport path, while an additional cycle-consistency constraint encourages the two maps to be approximate inverses. The resulting objective can be trained using standard neural networks and standard stochastic optimization.

Our main contributions are:

(1) **A target-free learning framework for OT maps.** We introduce synchronized forward-backward interpolants that provide intermediate-time supervision without requiring paired samples, precomputed couplings, or externally specified trajectories.

(2) **Provable recovery of the quadratic OT map.** Under suitable assumptions, we show that minimizing our objective recovers the population quadratic OT map and its inverse.

(3) **Scalable training with standard neural architectures.** Our method avoids iterative Reflow-style retraining (Xingchao Liu et al., 2023), minibatch OT approximations (Pooladian et al., 2023; Tong et al., 2023), and Input Convex Neural Networks (Amos et al., 2017), enabling efficient optimization with expressive architectures such as ResNets or Transformers.

(4) **Empirical validation across generative and scientific benchmarks.** Across synthetic, image, and scientific datasets, our approach yields stable training, accurate one-step generation, and meaningful interpolations competitive with representative OT-based and flow-based generative methods.

At the theoretical level, our objective is motivated by an L^2 projection view of Wasserstein barycenters under independence constraints (S. Xu et al., 2023). This perspective suggests that enforcing independence-like intermediate agreement between map-induced interpolants can identify the quadratic OT structure without explicitly solving for a coupling.

3.2.2. Related Work.

3.2.2.1. *Neural Optimal Transport and Direct Map Learning.* Classical OT provides a rigorous variational formulation for matching probability measures (Villani, 2009; Santambrogio, 2015; Peyré et al., 2019), and the quadratic-cost case admits the Brenier-map structure (Brenier, 1991). However, computing OT maps from samples in high dimension remains challenging. Entropic solvers such as Sinkhorn iterations (Cuturi, 2013) are effective for discrete OT but do not directly yield parametric out-of-sample maps. Neural OT methods address this by parameterizing transport maps, costs, or Kantorovich potentials with neural networks (Korotin et al., 2023). A prominent Brenier-inspired line represents maps as gradients of convex potentials, often implemented using Input Convex Neural Networks (Amos et al., 2017; Makkuva et al., 2020). While theoretically appealing, such convex-potential parameterizations can restrict architectural flexibility and complicate optimization. Our method instead learns forward and backward maps using standard neural architectures and enforces OT structure through intermediate-time map-induced matching and cycle consistency.

3.2.2.2. *Flow Matching and Coupling Design.* Flow Matching trains continuous normalizing flows by regressing velocity fields along prescribed probability paths (Lipman et al., 2022). The geometry of these paths depends critically on the coupling between source and target samples.

Rectified Flow and Reflow aim to learn straighter trajectories, but often require repeated simulation or retraining stages (Xingchao Liu et al., 2023). Minibatch OT and multisample couplings improve trajectory straightness by matching source and target samples within each batch (Pooladian et al., 2023; Tong et al., 2023), but introduce batch-dependent approximations and additional OT computation. Recent semi-discrete and structured coupling methods further improve scalability or few-step generation by replacing batch OT with semi-discrete or hierarchical coupling mechanisms (Mousavi-Hosseini et al., 2025; Kong et al., 2026; Yichi Zhang et al., 2025; Ma et al., 2026). In contrast, our method does not require precomputed couplings, minibatch OT, or semi-discrete pairing. Instead, it creates target-free intermediate supervision by matching the forward and backward interpolants induced by the learned maps.

3.2.2.3. *Single-Step Generative Models.* Accelerating generative sampling to one or a few steps is a major goal in diffusion and flow-based modeling. Consistency models and consistency trajectory models learn self-consistent maps along probability-flow trajectories (Y. Song et al., 2023; D. Kim et al., 2023), while progressive distillation and shortcut models reduce the number of sampling steps by distilling or shortcutting iterative samplers (Salimans et al., 2022; Frans et al., 2024). Recent mean-flow and flow-map-matching approaches further seek direct map-based generation (Geng et al., 2025; Boffi et al., 2025). These approaches primarily focus on sampling acceleration, often through distillation, consistency constraints, or specialized schedules. Our goal is different: we seek to recover the quadratic OT map and its associated displacement interpolation from unpaired samples, thereby obtaining one-step generation as a consequence of OT map learning.

3.2.2.4. *Unpaired Translation and Cycle Consistency.* Cycle-consistent adversarial methods learn mappings between unpaired domains by combining adversarial distribution matching with inverse-consistency constraints (Zhu et al., 2017). These methods are highly influential in image translation, but they do not in general identify the quadratic OT map or the Wasserstein geodesic between distributions. Our use of cycle consistency is instead tied to OT map recovery: the forward and backward maps define synchronized displacement interpolants, and intermediate matching provides geometric supervision along the transport path.

3.3. Background and Problem Setup

3.3.1. Quadratic Optimal Transport and McCann Interpolation. Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures with finite second moments, where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$ with finite second moment. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, and we consider the quadratic transport cost

$$k(x, y) = \frac{1}{2}\|x - y\|^2.$$

The Monge formulation of quadratic optimal transport seeks a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes μ_0 forward to μ_1 while minimizing the expected transport cost:

$$\inf_{T_{\#}\mu_0=\mu_1} \mathcal{K}(T) := \inf_{T_{\#}\mu_0=\mu_1} \mathbb{E}_{X_0\sim\mu_0}\left[\frac{1}{2}\|T(X_0) - X_0\|^2\right].$$

Here $T_{\#}\mu_0 = \mu_1$ means that if $X_0 \sim \mu_0$, then $T(X_0) \sim \mu_1$.

Under standard regularity assumptions, in particular when μ_0 is absolutely continuous with respect to the Lebesgue measure, Brenier's theorem states that the quadratic OT problem admits a unique optimal Monge map T^*, up to μ_0-almost everywhere equivalence. Moreover, this map has the form

$$T^* = \nabla\phi,$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a convex potential. Therefore, in the quadratic case, the optimal transport problem admits a canonical deterministic solution. In particular, $T^*_{\#}\mu_0 = \mu_1$, and the optimal Monge cost satisfies

$$\mathcal{K}(T^*) = \frac{1}{2}W_2^2(\mu_0, \mu_1),$$

where W_2 denotes the Wasserstein-2 distance.

The optimal map T^* induces a canonical interpolation between the endpoint distributions. For $t \in [0, 1]$, the McCann displacement interpolation is defined by

$$X_t = (1 - t)X_0 + tT^*(X_0), \qquad X_0 \sim \mu_0, \tag{3.1}$$

with corresponding marginal law

$$\mu_t = \mathcal{L}(X_t) = \big((1-t)\mathrm{Id} + tT^*\big)_{\#}\mu_0.$$

3.3.2. Dynamic Optimal Transport Formulation: The Benamou-Brenier Formulation. The McCann interpolation also has a dynamic interpretation. Along the path induced by T^*, each particle travels on a straight line with constant velocity

$$v^*(X_0) = T^*(X_0) - X_0.$$

Equivalently,

$$\frac{d}{dt}X_t = T^*(X_0) - X_0.$$

Therefore, the kinetic action of this path is

$$\int_0^1 \mathbb{E}\left[\frac{1}{2}\left\|\frac{d}{dt}X_t\right\|^2\right] dt = \mathbb{E}_{X_0\sim\mu_0}\left[\frac{1}{2}\|T^*(X_0) - X_0\|^2\right] = \frac{1}{2}W_2^2(\mu_0,\mu_1).$$

This dynamic interpretation is consistent with the Benamou–Brenier formulation of quadratic OT, and it can be written as

$$\frac{1}{2}W_2^2(\mu_0,\mu_1) = \inf_{\rho_t, v_t}\int_0^1\int_{\mathbb{R}^d}\frac{1}{2}\|v_t(x)\|^2\, d\rho_t(x)\, dt,$$

where the infimum is taken over distributional paths $(\rho_t)_{t\in[0,1]}$ and velocity fields $(v_t)_{t\in[0,1]}$ satisfying the continuity equation

$$\partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0, \qquad \rho_0 = \mu_0, \quad \rho_1 = \mu_1,$$

The continuity equation is understood in the distributional sense: for every test function $\varphi \in C_c^\infty((0,1)\times\mathbb{R}^d)$, the pair (ρ_t, v_t) satisfies

$$\int_0^1\int_{\mathbb{R}^d}\left(\partial_t\varphi(t,x) + \langle\nabla\varphi(t,x), v_t(x)\rangle\right)\rho_t(dx)\, dt = 0.$$

Thus, the McCann interpolation induced by T^* is not merely an endpoint-matching construction; it realizes the minimum-action path in the Benamou–Brenier dynamic formulation.

3.3.3. Flow Matching and Rectified Flow. Flow matching learns a time-dependent vector field whose induced probability flow transports a source distribution to a target distribution. Specifically, one considers the ODE

$$\frac{dZ_t}{dt} = v_\theta(Z_t, t), \qquad Z_0 \sim \mu_0,$$

and trains v_θ using a prescribed reference interpolation between samples from μ_0 and μ_1. A common choice is the linear interpolation

$$X_t = (1-t)X_0 + tX_1, \qquad X_0 \sim \mu_0, \quad X_1 \sim \mu_1,$$

where (X_0, X_1) is sampled from a chosen coupling, often the independent coupling. The corresponding reference velocity is

$$u(X_0, X_1) = X_1 - X_0.$$

Flow matching then minimizes the regression objective

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \int_0^1 \mathbb{E}\left[\|X_1 - X_0 - v_\theta(X_t, t)\|^2\right] dt.$$

At the population level, the minimizer over all measurable vector fields is

$$v^*(z,t) = \mathbb{E}\left[X_1 - X_0 \mid X_t = z\right].$$

The reference marginals $\mu_t = \mathcal{L}(X_t)$ then satisfy the continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_T^*) = 0$$

in the weak sense. If this continuity equation is well posed for v^*, then the ODE driven by v^* has the same time marginals as the reference interpolation.

Rectified flow applies this projection iteratively. After learning a velocity field, it constructs new endpoint pairings from the learned flow and retrains on the resulting trajectories. This reflow procedure preserves the endpoint marginals while progressively straightening the learned paths.

In contrast, our framework does not learn a time-dependent velocity field or repeatedly simulate an ODE. Instead, we learn endpoint transport maps directly and use their induced McCann-type interpolants to impose intermediate-time consistency.

3.3.4. Independence constraints and Wasserstein barycenters. Following S. Xu et al. (2023), we recall the connection between independence-constrained L^2 projection and Wasserstein barycenters. Let $Z \in \{0, 1\}$ be a group label with

$$\mathbb{P}(Z = k) = \lambda_k, \qquad \mu_k = \mathcal{L}(X \mid Z = k), \qquad k \in \{0, 1\}.$$

Consider the independence-constrained projection problem

$$\inf_{\overline{X}: \overline{X} \perp Z} \mathbb{E}\left[\|X - \overline{X}\|^2\right].$$

This constraint means that the conditional law of $\overline{X}$ is the same for both groups. Equivalently, there exists a common distribution ν such that

$$\mathcal{L}(\overline{X} \mid Z = 0) = \mathcal{L}(\overline{X} \mid Z = 1) = \nu.$$

For a fixed common law ν, the optimal way to map the k-th group distribution μ_k to ν under quadratic distortion has cost

$$W_2^2(\mu_k, \nu).$$

Therefore, the independence-constrained projection problem reduces to choosing the best common law:

$$\inf_{\overline{X}: \overline{X} \perp Z} \mathbb{E}\|X - \overline{X}\|^2 = \inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=0}^{1} \lambda_k W_2^2(\mu_k, \nu).$$

Any minimizer ν^* of the right-hand side is the quadratic Wasserstein barycenter of the group-conditional distributions (S. Xu et al., 2023).

This result shows that an independence constraint can convert an L^2 projection problem into a Wasserstein barycenter problem: independence enforces a shared output law, while the quadratic projection cost measures the optimal transport cost from each group distribution to that shared law.

3.4. Proposed Framework

3.4.1. Motivation and Approach. Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures with finite second moments. We aim to learn the quadratic optimal transport map

$$T^* : \mu_0 \to \mu_1$$

from unpaired samples, together with its induced displacement interpolation. Rather than computing a coupling or imposing a convex-potential parameterization, we learn transport maps directly using standard neural architectures.

Our starting point is the McCann interpolation induced by the optimal map:

$$\mu_t = \big((1-t)\mathrm{Id} + tT^*\big)_{\#}\mu_0, \qquad t \in [0,1].$$

This path captures not only endpoint marginal matching, but also the intermediate trajectory laws generated by the transport map. Thus, a learned map should be regularized through its induced path, not only through its endpoint distribution.

Flow matching and rectified flow learn such dynamics by fitting a time-dependent velocity field along a prescribed interpolation (Xingchao Liu et al., 2023; Q. Liu et al., 2024). Some recent variants incorporate additional OT or Schrödinger Bridge structure through specialized objectives or parameterizations (Kornilov et al., 2024; Gushchin et al., 2024). In contrast, our method remains map-based: it learns a forward map $T : \mu_0 \to \mu_1$ and a backward map $S : \mu_1 \to \mu_0$. These maps induce two McCann-type interpolants moving from opposite directions:

$$X_t^F = (1-t)X_0 + tT(X_0), \qquad X_0 \sim \mu_0,$$

and

$$X_t^B = (1-t)S(X_1) + tX_1, \qquad X_1 \sim \mu_1.$$

If $T = T^*$ and $S = (T^*)^{-1}$ on the relevant supports, then both interpolants recover the same displacement path:

$$\mathcal{L}(X_t^F) = \mathcal{L}(X_t^B), \qquad t \in [0,1].$$

This motivates our midpoint consistency principle: at intermediate times, the forward and backward interpolants should have matching laws. Equivalently, if a direction label $Z \in \{0, 1\}$ indicates whether an intermediate sample comes from the forward or backward interpolant, then midpoint consistency enforces an independence-type condition

$$X_t \perp Z,$$

which is equivalent to

$$\mathcal{L}(X_t^F) = \mathcal{L}(X_t^B).$$

This connects our construction to the independence-constrained L^2 projection viewpoint of Wasserstein barycenters (S. Xu et al., 2023): independence enforces a shared law, while the quadratic transport cost controls the displacement toward that law.

Our objective combines quadratic transport cost, adversarial endpoint and intermediate-time marginal matching, and cycle consistency. Together, these terms define a bidirectional, map-based framework for learning OT trajectories from unpaired samples.

Now, we introduce a symmetric supervision framework that regularizes the learning of the transport map by requiring consistency between forward and backward interpolants at intermediate time steps.

3.4.2. Parametric Maps and Synchronized Interpolants. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and $\mathcal{G} = \{g_\phi : \phi \in \Phi\}$ be parametric families of measurable maps (e.g., deep neural networks) from $\mathcal{X}$ to $\mathcal{X}$. We interpret f_θ as the forward transport map ($\mu_0 \to \mu_1$) and g_ϕ as its approximate inverse ($\mu_1 \to \mu_0$).

Following the McCann interpolation defined in (3.1), we construct synchronized forward and backward trajectories. For any $t \in [0, 1]$, we define:

- Forward interpolant: $F_t^\theta(x) := (1 - t)x + tf_\theta(x)$, inducing $\mu_t^\theta := (F_t^\theta)_\# \mu_0$.
- Backward interpolant: $G_{1-t}^\phi(y) := ty + (1 - t)g_\phi(y)$, inducing $\nu_t^\phi := (G_{1-t}^\phi)_\# \mu_1$.

Note that at $t = 0$, $\mu_0^\theta = \mu_0$, and at $t = 1$, $\nu_1^\phi = \mu_1$. The goal is to ensure that for all $t \in (0, 1)$, the distributions μ_t^θ and ν_t^ϕ coincide, effectively pinning the learned path to the unique W_2 geodesic.

3.4.3. Variational Distribution Matching via Adversarial Detection. To enforce the intermediate consistency constraint $\mu_t^\theta = \nu_t^\phi$ without requiring explicit density estimation, we cast distribution matching as a binary detection problem. This allows us to quantify the discrepancy between distributions using a variational mutual-information objective.

Let $P, Q \in \mathcal{P}(\mathcal{X})$ be the two distributions to be matched (in our context, $P = \mu_t^\theta$ and $Q = \nu_t^\phi$). We introduce a binary latent variable $Z \in \{0, 1\}$ with a balanced prior $\mathbb{P}(Z = 1) = \mathbb{P}(Z = 0) = \frac{1}{2}$. We then draw an observation $X \in \mathcal{X}$ conditionally on Z:

$$X \mid (Z = 1) \sim P, \qquad X \mid (Z = 0) \sim Q.$$

Let $\pi := \frac{1}{2}(P + Q)$ denote the marginal mixture distribution of X. Under this model, the distributions P and Q are identical if and only if the observation X provides no information about the label Z.

The degree of "distinguishability" between P and Q is precisely captured by the mutual information $I(Z; X)$, defined as:

$$I(Z; X) = H(Z) - H(Z \mid X) = \log 2 - H(Z \mid X), \tag{3.2}$$

where H denotes the Shannon entropy. Since $I(Z; X)$ is bounded in $[0, \log 2]$, it serves as a robust discrepancy measure: $I(Z; X) = 0$ if and only if Z and X are independent, which occurs if and only if $P = Q$ π-a.e.

While the posterior $p(Z \mid X)$ is generally unknown, we can bound the mutual information from below using a class of parametric discriminators $\mathcal{D} = \{D_\omega : \mathcal{X} \to (0, 1)\}$, where $D_\omega(x)$ approximates the posterior probability $\mathbb{P}(Z = 1 \mid X = x)$. We define the binary cross-entropy (BCE) functional:

$$\mathcal{J}_\mathcal{D}(P, Q) := \sup_{D \in \mathcal{D}} \Big\{ \mathbb{E}_{X \sim P}[\log D(X)] + \mathbb{E}_{X \sim Q}[\log(1 - D(X))] \Big\}. \tag{3.3}$$

In full generality where $\mathcal{D}$ is the set of all measurable functions into $(0, 1)$, the supremum is attained by the Bayes-optimal detector:

$$D^*(x) = \mathbb{P}(Z = 1 \mid X = x) = \frac{p(x)}{p(x) + q(x)}, \tag{3.4}$$

where p, q are the densities of P, Q with respect to the mixture π. Substituting D^* into (3.3) reveals the fundamental identity:

$$I(Z;X) = \log 2 + \frac{1}{2}\mathcal{J}_{\text{all}}(P,Q). \tag{3.5}$$

Equation 3.5 establishes that maximizing the detection objective is equivalent to estimating the Jensen-Shannon Divergence (JSD). Specifically:

$$I(Z;X) = \text{JSD}(P,Q) = \frac{1}{2}\text{KL}(P\|\pi) + \frac{1}{2}\text{KL}(Q\|\pi). \tag{3.6}$$

In practice, when $\mathcal{D}$ is a restricted class (e.g., neural networks), (3.3) provides an adversarially-estimated JSD. This serves as a differentiable proxy for distribution matching, where the transport maps f_θ and g_ϕ act as generators attempting to minimize the ability of D to distinguish between the forward and backward interpolants.

3.4.4. The Multi-Objective Population Risk. We define the population objective $\mathcal{F}_\lambda(f,g)$ as a weighted combination of transport cost, distributional consistency, and algebraic invertibility:

$$\mathcal{F}_\lambda(f,g) := \lambda\mathcal{A}(f,g) + \mathcal{B}_{\text{JSD}}(f,g) + \mathcal{C}(f,g), \tag{3.7}$$

for $\lambda > 0$. Here, the constituent terms are:

- Kinetic Energy :

$$\mathcal{A}(f,g) := \frac{1}{2}\int_{\mathcal{X}} \|f(x) - x\|^2 \, d\mu_0(x) + \frac{1}{2}\int_{\mathcal{X}} \|g(y) - y\|^2 \, d\mu_1(y),$$

  This term encourages f and g to minimize the L^2 transport cost.

- Boundary Matching:

$$\mathcal{B}_{\text{JSD}}(f,g) := \text{JSD}(f_\#\mu_0, \mu_1) + \text{JSD}(g_\#\mu_1, \mu_0).$$

  This provides two-marginal distribution matching supervision.

- Cycle Consistency:

$$\mathcal{C}(f,g) := \int_{\mathcal{X}} \|g(f(x)) - x\|^2 \, d\mu_0(x) + \int_{\mathcal{X}} \|f(g(y)) - y\|^2 \, d\mu_1(y).$$

This enforces that f and g are closely inverse to each other.

Now, we show that under the universality representation capability of deep neural networks, the proposed multi-objective population risk guarantees the global minimizer is close to the true OT map.

THEOREM 3.4.1 (Guaranteed OT Recovery under Universality Approximation and Annealing Transport). *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact with diameter $D_{\mathcal{X}} := \operatorname{diam}(\mathcal{X})$, and let $\mu_0, \mu_1 \in \mathcal{P}_{2,ac}(\mathcal{X})$. Let $T^* : \mu_0 \to \mu_1$ and $S^* : \mu_1 \to \mu_0$ be the unique quadratic Brenier maps, and let $W := W_2(\mu_0, \mu_1) > 0$. Assume that the population admissible generator and adversarial classes are expressive enough to approximate the Brenier pair (T^*, S^*) and the JSD minimizer. Then, for every $\varepsilon > 0$, there exists a constant*

$$\alpha_\varepsilon = \alpha(\varepsilon, \mu_0, \mu_1, D_{\mathcal{X}}) > 0$$

such that if

$$\lambda \leq \frac{\alpha_\varepsilon}{W^2},$$

then

$$\|f_\lambda - T^*\|_{L^2(\mu_0)} + \|g_\lambda - S^*\|_{L^2(\mu_1)} < \varepsilon.$$

PROOF. Since (T^*, S^*) is admissible,

$$(T^*)_{\#}\mu_0 = \mu_1, \qquad (S^*)_{\#}\mu_1 = \mu_0.$$

Moreover, the forward and reverse Brenier maps are inverse to each other almost everywhere:

$$S^* \circ T^* = \mathrm{Id}_{\mu_0}, \qquad T^* \circ S^* = \mathrm{Id}_{\mu_1}.$$

Therefore,

$$\mathcal{B}_{\mathrm{JSD}}(T^*, S^*) = 0, \qquad \mathcal{C}(T^*, S^*) = 0.$$

By the definition of $\mathcal{A}$,

$$\mathcal{A}(T^*, S^*) = W^2.$$

Hence, by global optimality,

$$\mathcal{F}_\lambda(f_\lambda, g_\lambda) \le \mathcal{F}_\lambda(T^*, S^*) = \lambda W^2.$$

Since $\mathcal{A}$, $\mathcal{B}_{\mathrm{JSD}}$, and $\mathcal{C}$ are nonnegative, it follows that

$$\mathcal{A}(f_\lambda, g_\lambda) \le W^2,$$

$$\mathrm{JSD}((f_\lambda)_{\#}\mu_0, \mu_1) \le \lambda W^2, \qquad \mathrm{JSD}((g_\lambda)_{\#}\mu_1, \mu_0) \le \lambda W^2,$$

and

$$\mathcal{C}(f_\lambda, g_\lambda) \le \lambda W^2.$$

We next convert the JSD bounds into W_2 bounds. By Pinsker's inequality for JSD,

$$\mathrm{TV}(P, Q) \le \sqrt{2\,\mathrm{JSD}(P, Q)}.$$

Since $\mathcal{X}$ has diameter $D_{\mathcal{X}}$,

$$W_2(P, Q) \le D_{\mathcal{X}}\,\mathrm{TV}(P, Q)^{1/2} \le D_{\mathcal{X}}\big(2\,\mathrm{JSD}(P, Q)\big)^{1/4}.$$

Therefore,

$$W_2((f_\lambda)_{\#}\mu_0, \mu_1) \le D_{\mathcal{X}}(2\lambda W^2)^{1/4},$$

and similarly,

$$W_2((g_\lambda)_{\#}\mu_1, \mu_0) \le D_{\mathcal{X}}(2\lambda W^2)^{1/4}.$$

Define the individual transport costs

$$I_f := \int_{\mathcal{X}} \|f_\lambda(x) - x\|^2 \, d\mu_0(x), \qquad I_g := \int_{\mathcal{X}} \|g_\lambda(y) - y\|^2 \, d\mu_1(y).$$

Then

$$\frac{1}{2} I_f + \frac{1}{2} I_g = \mathcal{A}(f_\lambda, g_\lambda) \le W^2.$$

Since the graph coupling induced by f_λ is an admissible coupling between μ_0 and $(f_\lambda)_\# \mu_0$,

$$W_2(\mu_0, (f_\lambda)_\# \mu_0) \leq I_f^{1/2}.$$

Thus, by the triangle inequality,

$$I_f^{1/2} \geq W - W_2((f_\lambda)_\# \mu_0, \mu_1).$$

Similarly,

$$I_g^{1/2} \geq W - W_2((g_\lambda)_\# \mu_1, \mu_0).$$

We now claim that, for every $\tau > 0$, there exist constants $\eta_\tau > 0$ and $\delta_\tau > 0$ such that every measurable map $h : \mathcal{X} \to \mathcal{X}$ satisfying

$$W_2(h_\# \mu_0, \mu_1) \leq \eta_\tau$$

and

$$\int_{\mathcal{X}} \|h(x) - x\|^2 \, d\mu_0(x) \leq W^2 + \delta_\tau$$

also satisfies

$$\|h - T^*\|_{L^2(\mu_0)} < \tau.$$

We then take $\eta_{\frac{\varepsilon}{2}}, \delta_{\frac{\varepsilon}{2}}$ small enough so that both the forward and reverse stability statements hold with tolerance $\tau = \varepsilon/2$. Now define

$$a_\varepsilon := \min\left\{\eta_{\frac{\varepsilon}{2}}, \frac{\delta_{\frac{\varepsilon}{2}}}{2W}, \frac{W}{2}\right\}, \qquad \alpha_\varepsilon := \frac{1}{2}\left(\frac{a_\varepsilon}{D_{\mathcal{X}}}\right)^4.$$

If $\lambda \leq \frac{\alpha_\varepsilon}{W^2}$, then

$$D_{\mathcal{X}}(2\lambda W^2)^{1/4} \leq a_\varepsilon.$$

Therefore,

$$W_2((f_\lambda)_\# \mu_0, \mu_1) \leq a_\varepsilon \leq \eta_\varepsilon,$$

and

$$W_2((g_\lambda)_\# \mu_1, \mu_0) \leq a_\varepsilon \leq \eta_\varepsilon.$$

Using the lower bound for I_g,

$$I_g \geq (W - a_\varepsilon)^2.$$

Together with

$$\frac{1}{2} I_f + \frac{1}{2} I_g \leq W^2,$$

this gives

$$I_f \leq 2W^2 - (W - a_\varepsilon)^2 = W^2 + 2W a_\varepsilon - a_\varepsilon^2 \leq W^2 + \delta_\varepsilon,$$

because $a_\varepsilon \leq \delta_\varepsilon/(2W)$. The same argument gives

$$I_g \leq W^2 + \delta_\varepsilon.$$

The forward and reverse stability statements now imply

$$\|f_\lambda - T^*\|_{L^2(\mu_0)} < \varepsilon/2, \qquad \|g_\lambda - S^*\|_{L^2(\mu_1)} < \varepsilon/2.$$

Summing the two estimates yields

$$\|f_\lambda - T^*\|_{L^2(\mu_0)} + \|g_\lambda - S^*\|_{L^2(\mu_1)} < \varepsilon.$$

It remains to prove the claim. We assume by contradiction that there exists $\tau_0 > 0$ such that, for every $n \in \mathbb{N}$, one can find a measurable map $h_n : \mathcal{X} \to \mathcal{X}$ with

$$W_2((h_n)_\# \mu_0, \mu_1) \leq \frac{1}{n},$$

$$\int_{\mathcal{X}} \|h_n(x) - x\|^2 \, d\mu_0(x) \leq W^2 + \frac{1}{n},$$

but

$$\|h_n - T^*\|_{L^2(\mu_0)} \geq \tau_0.$$

The marginal convergence also gives a lower bound on the map costs. Indeed, by the triangle inequality,

$$W \leq W_2(\mu_0, (h_n)_\# \mu_0) + W_2((h_n)_\# \mu_0, \mu_1).$$

Since the graph coupling induced by h_n is admissible,

$$W_2(\mu_0, (h_n)_{\#}\mu_0) \leq \left(\int_{\mathcal{X}} \|h_n(x) - x\|^2 \, d\mu_0(x) \right)^{1/2}.$$

Hence

$$\left(\int_{\mathcal{X}} \|h_n(x) - x\|^2 \, d\mu_0(x) \right)^{1/2} \geq W - \frac{1}{n}.$$

For all sufficiently large n,

$$\left(W - \frac{1}{n} \right)^2 \leq \int_{\mathcal{X}} \|h_n(x) - x\|^2 \, d\mu_0(x) \leq W^2 + \frac{1}{n}.$$

Therefore,

$$\int_{\mathcal{X}} \|h_n(x) - x\|^2 \, d\mu_0(x) \longrightarrow W^2.$$

Define the graph couplings

$$\gamma_n := (\mathrm{Id}, h_n)_{\#}\mu_0.$$

Since $\mathcal{X} \times \mathcal{X}$ is compact, a subsequence, not relabeled, converges weakly to some probability measure γ. The first marginal of γ is μ_0. The second marginal is μ_1, because $(h_n)_{\#}\mu_0 \to \mu_1$ in W_2. Thus

$$\gamma \in \Pi(\mu_0, \mu_1).$$

Moreover, since $(x, y) \mapsto \|x - y\|^2$ is continuous and bounded on $\mathcal{X} \times \mathcal{X}$,

$$\int \|x - y\|^2 \, d\gamma(x, y) = \lim_{n \to \infty} \int \|x - y\|^2 \, d\gamma_n(x, y) = W^2.$$

Thus γ is an optimal quadratic coupling between μ_0 and μ_1. Since μ_0 is absolutely continuous, this optimal coupling is unique and is induced by the Brenier map:

$$\gamma = (\mathrm{Id}, T^*)_{\#}\mu_0.$$

Consequently, along the chosen subsequence,

$$(\mathrm{Id}, h_n)_{\#}\mu_0 \Rightarrow (\mathrm{Id}, T^*)_{\#}\mu_0 \qquad \text{in } \mathcal{P}(\mathcal{X} \times \mathcal{X}).$$

It remains to convert this weak convergence of graph couplings into $L^2(\mu_0)$-convergence of the maps. Fix $\rho > 0$. Since T^* is measurable and μ_0 is finite, Lusin's theorem gives a compact set $K_\rho \subset \mathcal{X}$ such that

$$\mu_0(\mathcal{X} \setminus K_\rho) < \rho$$

and $T^*|_{K_\rho}$ is continuous. Since $\mathcal{X}$ is compact, there exists $R < \infty$ such that

$$\|x\| \leq R \qquad \text{for all } x \in \mathcal{X}.$$

Because $T^*(K_\rho) \subset \mathcal{X}$, the bounded Tietze extension theorem, applied coordinate-wise, gives a continuous map $\widetilde{T}_\rho : \mathcal{X} \to \mathbb{R}^d$ such that

$$\widetilde{T}_\rho = T^* \qquad \text{on } K_\rho,$$

and

$$\|\widetilde{T}_\rho(x)\| \leq C_{\mathcal{X}} \qquad \text{for all } x \in \mathcal{X},$$

where $C_{\mathcal{X}} < \infty$ depends only on $\mathcal{X}$, not on ρ. Since $h_n(x) \in \mathcal{X}$, $T^*(x) \in \mathcal{X}$, and $\widetilde{T}_\rho$ is uniformly bounded, there exists $M_{\mathcal{X}} < \infty$, independent of n and ρ, such that

$$\|h_n(x) - \widetilde{T}_\rho(x)\|^2 + \|T^*(x) - \widetilde{T}_\rho(x)\|^2 \leq M_{\mathcal{X}}$$

for all $x \in \mathcal{X}$. The function

$$(x, y) \mapsto \|y - \widetilde{T}_\rho(x)\|^2$$

is bounded and continuous on $\mathcal{X} \times \mathcal{X}$. Therefore, weak convergence of γ_n to $(\mathrm{Id}, T^*)_\# \mu_0$ gives

$$\int_{\mathcal{X}} \|h_n(x) - \widetilde{T}_\rho(x)\|^2 \, d\mu_0(x) \longrightarrow \int_{\mathcal{X}} \|T^*(x) - \widetilde{T}_\rho(x)\|^2 \, d\mu_0(x).$$

Because $\widetilde{T}_\rho = T^*$ on K_ρ, the right-hand side is bounded by

$$M_{\mathcal{X}} \mu_0(\mathcal{X} \setminus K_\rho) \leq M_{\mathcal{X}} \rho.$$

Hence

$$\limsup_{n \to \infty} \|h_n - \widetilde{T}_\rho\|^2_{L^2(\mu_0)} \leq M_{\mathcal{X}} \rho,$$

and also

$$\|\widetilde{T}_\rho - T^*\|^2_{L^2(\mu_0)} \leq M_{\mathcal{X}}\rho.$$

Using

$$\|h_n - T^*\|^2 \leq 2\|h_n - \widetilde{T}_\rho\|^2 + 2\|\widetilde{T}_\rho - T^*\|^2,$$

we obtain

$$\limsup_{n\to\infty} \|h_n - T^*\|^2_{L^2(\mu_0)} \leq 4M_{\mathcal{X}}\rho.$$

Since $\rho > 0$ is arbitrary,

$$\|h_n - T^*\|_{L^2(\mu_0)} \longrightarrow 0$$

along the chosen subsequence. This contradicts

$$\|h_n - T^*\|_{L^2(\mu_0)} \geq \tau_0.$$

Thus, the forward stability claim holds. Applying the same argument to the reverse OT problem gives the corresponding constants for S^*. That completes the proof. □

3.4.5. Training Procedure. To minimize the population objective in (3.7) in practice, we parametrize the forward and backward maps T and S by neural networks $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ and $g_\phi : \mathbb{R}^d \to \mathbb{R}^d$, respectively. We also introduce discriminator networks $D^0_{\omega_0}$, $D^1_{\omega_1}$, and $D^t_{\omega_t}$ to enforce distributional matching at times 0, 1, and $t \in (0, 1)$, respectively. The intermediate-time discriminator $D^t_{\omega_t}$ is optional. In practice, we use an epoch-dependent annealed weight for the transport regularization term. Let k denote the current epoch and let λ_k be the corresponding transport weight. Specifically, we use an exponential decay schedule

$$\lambda_k = \lambda_0^k, \qquad 0 < \lambda_0 < 1.$$

This gradually decreases the influence of the transport regularization during training. We summarize the resulting training procedure in Algorithm 2, which is provided in Appendix B.2.

3.5. Theoretical Motivation Underlying the Forward-Backward Supervision

This section establishes the theoretical foundations of the proposed forward–backward supervision framework. The core structural results demonstrate that the joint objective operates via two complementary mechanisms:

(1) **Feasible-Set Reduction:** While endpoint matching alone leaves the transport maps highly under-determined, cycle consistency restricts the solution space by enforcing approximate invertibility at a controlled spatial resolution.
(2) **Optimization-Level Certification:** Among all marginal-matching maps, the L^2-loss minimization uniquely identifies the true OT pair. At the optimization level, persistent low cycle loss before and after constrained transport-gradient updates certifies proximity to a cycle-stationary set. Unless the forward and backward updates are precisely coordinated by an inverse-tangency cancellation, this corresponds locally to low constrained transport gradient.

Additionally, we show that cycle consistency structurally suppresses map-level collapse by strictly bounding the conditional variance of preimages.

3.5.0.1. *Standing assumptions and notation.* Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact, convex domain, and let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathcal{X})$ be absolutely continuous measures. We assume that the unique quadratic OT map $T^* : \mu_0 \to \mu_1$ exists with a μ_0-a.e. unique inverse $S^* : \mu_1 \to \mu_0$ on their respective supports.

Given a forward map f and a backward map g, the map-induced McCann interpolants and their pushed-forward measures are defined for $t \in [0, 1]$ by:

$$F_t^f = (1-t)\mathrm{Id} + tf, \qquad \mu_t^f := (F_t^f)_{\#}\mu_0, \tag{3.8}$$

$$G_{1-t}^g = t\mathrm{Id} + (1-t)g, \qquad \nu_t^g := (G_{1-t}^g)_{\#}\mu_1. \tag{3.9}$$

We define the exact marginal-matching feasible set as:

$$\mathcal{M}_{\mathrm{match}} := \{(f, g) : f_{\#}\mu_0 = \mu_1,\ g_{\#}\mu_1 = \mu_0\}. \tag{3.10}$$

The bidirectional quadratic action $\mathcal{A}(f,g)$ and the two-sided cycle loss $\mathcal{C}(f,g)$ are defined as:

$$\mathcal{A}(f,g) := \frac{1}{2}\mathbb{E}_{X\sim\mu_0}\|f(X)-X\|^2 + \frac{1}{2}\mathbb{E}_{Y\sim\mu_1}\|g(Y)-Y\|^2, \tag{3.11}$$

$$\mathcal{C}(f,g) := \mathbb{E}_{X\sim\mu_0}\|g(f(X))-X\|^2 + \mathbb{E}_{Y\sim\mu_1}\|f(g(Y))-Y\|^2. \tag{3.12}$$

We compactly denote the joint map pair by $z := (f,g)$ and the optimal pair by $z^* := (T^*, S^*)$, equipped with the product L^2-metric:

$$\|z-z^*\|^2 := \|f-T^*\|^2_{L^2(\mu_0)} + \|g-S^*\|^2_{L^2(\mu_1)}. \tag{3.13}$$

3.5.1. Feasible Class Reduction via Cycle Consistency. Learning optimal transport maps is fundamentally challenged by the under-determination of the coupling class $\Pi(\mu_0,\mu_1)$. Any coupling $\gamma \in \Pi(\mu_0,\mu_1)$ can be disintegrated as $\gamma(dx,dy) = \mu_0(dx)K(dy \mid x)$, where a stochastic kernel K introduces output randomization (one-to-many matching). While a deterministic forward map f eliminates this stochasticity via a graph-supported Monge coupling $\gamma_f := (\mathrm{Id}, f)_{\#}\mu_0$, it remains vulnerable to many-to-one collapse. The forward–backward pair (f,g) regularizes this residual degeneracy by leveraging cycle consistency to enforce approximate invertibility.

Crucially, endpoint matching alone is insufficient to identify the correct transport path; a map can match the target marginal perfectly while generating a highly suboptimal intermediate trajectory.

Proposition 3.5.1 (Incompatibility of endpoint matching). *Assume there exists a nontrivial μ_0-measure-preserving map $R : \mathcal{X} \to \mathcal{X}$ (i.e., $R_{\#}\mu_0 = \mu_0$ and $R \neq \mathrm{Id}$ on a set of positive μ_0-measure). Let T^* be the unique quadratic OT map from μ_0 to μ_1. Then the composed map*

$$f_R := T^* \circ R \tag{3.14}$$

satisfies the endpoint marginal condition $(f_R)_{\#}\mu_0 = \mu_1$, but induces non-identical intermediate marginals:

$$\mu_t^{f_R} \neq \mu_t^{T^*} \qquad \text{for some } t \in (0,1). \tag{3.15}$$

Proposition 3.5.1 underscores that endpoint constraints leave the transport path unconstrained. While classical OT resolves this ambiguity via cost minimization, deep generative models typically enforce marginal matching using soft adversarial or divergence penalties, leaving them susceptible to trajectory-incompatible solutions. The forward–backward construction introduces a complementary geometric regularizer: cycle consistency restricts the admissible map class to finite-resolution inverses.

DEFINITION 3.5.1 ((ε,η)-invertibility). *For $\varepsilon > 0$ and $\eta \in [0,1]$, the pair (f,g) is (ε,η)-invertible if*

$$\mathbb{P}_{X\sim\mu_0}\big(\|g(f(X)) - X\| > \varepsilon\big) + \mathbb{P}_{Y\sim\mu_1}\big(\|f(g(Y)) - Y\| > \varepsilon\big) \le \eta, \tag{3.16}$$

where ε defines the spatial evaluation resolution and η bounds the allowed failure probability.

PROPOSITION 3.5.2 (Resolution-bounded invertibility). *If $\mathcal{C}(f,g) \le \delta^2$, then for any $\varepsilon > 0$, the pair (f,g) is $\Big(\varepsilon, \min\{1, \frac{\delta^2}{\varepsilon^2}\}\Big)$-invertible.*

The above result shows that a small cycle loss guarantees, outside a failure set of measure at most δ^2/ε^2, the learned maps operate as exact inverses at spatial resolution ε. Consequently, cycle consistency structurally reduces the feasible space from general graph-supported, many-to-one transformations to an approximately invertible family. This constraint naturally complements the quadratic transport action: the action cost identifies low-energy displacements, while cycle consistency eliminates inverse-inconsistent paths.

3.5.2. Dynamic Stability of Low Cycle-loss Solutions. We now explain why non-OT inverse-compatible solutions are dynamically fragile under forward-backward training. Cycle consistency alone does not identify the unique OT map: a non-OT bijection f with $g = f^{-1}$ can achieve zero cycle loss. However, such a pair need not be stable under transport loss gradient updates constrained to preserve the endpoint marginals. Away from constrained stationary points of the transport loss, the forward and backward action corrections are nonzero. Here, we show that unless these corrections are coordinated to remain tangent to the inverse-consistency relations, such update would increase consistency loss.

Let

$$\mathcal{N}_0 := \{h \in L^2(\mu_0; \mathbb{R}^d) : h_{\#}\mu_0 = \mu_1\}, \qquad \mathcal{N}_1 := \{k \in L^2(\mu_1; \mathbb{R}^d) : k_{\#}\mu_1 = \mu_0\},$$

so that

$$\mathcal{M}_{\text{match}} = \mathcal{N}_0 \times \mathcal{N}_1.$$

Throughout this subsection, we assume that the constrained transport-gradient directions are well defined. Namely, for each $z = (f, g) \in \mathcal{M}_{\text{match}}$ considered below, let

$$\widetilde{U}_f := \Pi_{T_f\mathcal{N}_0}(f - \mathrm{Id}), \qquad \widetilde{U}_g := \Pi_{T_g\mathcal{N}_1}(g - \mathrm{Id}),$$

where $T_f\mathcal{N}_0$ and $T_g\mathcal{N}_1$ denote the feasible first-order perturbation spaces preserving the marginal constraints, and Π denotes L^2-orthogonal projection. Thus, $\widetilde{U}_f, \widetilde{U}_g$ are projected, not ambient, transport gradients. We consider one constrained projected-gradient step

$$f^+ = \mathrm{Retr}_f(-\alpha\widetilde{U}_f), \qquad g^+ = \mathrm{Retr}_g(-\alpha\widetilde{U}_g),$$

where the retractions return the tangent step to the marginal-matching sets:

$$(f^+)_{\#}\mu_0 = \mu_1, \qquad (g^+)_{\#}\mu_1 = \mu_0,$$

and

$$f^+ = f - \alpha\widetilde{U}_f + O(\alpha^2) \quad \text{in } L^2(\mu_0), \qquad g^+ = g - \alpha\widetilde{U}_g + O(\alpha^2) \quad \text{in } L^2(\mu_1).$$

Definition 3.5.2 (Projected Taylor regularity). *We say that $z = (f, g) \in \mathcal{M}_{\text{match}}$ satisfies projected Taylor regularity with constants (α_0, C_Q, C_R) if, for every $0 < \alpha \le \alpha_0$, the one-step cycle expansions hold in L^2:*

$$g^+(f^+(X)) - X = R_0(X) - \alpha\widetilde{W}_0(X) + \alpha^2\widetilde{Q}_0(X) + \mathcal{R}_0^\alpha(X), \tag{3.17}$$

$$f^+(g^+(Y)) - Y = R_1(Y) - \alpha\widetilde{W}_1(Y) + \alpha^2\widetilde{Q}_1(Y) + \mathcal{R}_1^\alpha(Y), \tag{3.18}$$

where

$$R_0(X) := g(f(X)) - X, \qquad R_1(Y) := f(g(Y)) - Y,$$

and

$$\widetilde{W}_0(X) := Dg(f(X))\widetilde{U}_f(X) + \widetilde{U}_g(f(X)),$$

$$\widetilde{W}_1(Y) := Df(g(Y))\widetilde{U}_g(Y) + \widetilde{U}_f(g(Y)).$$

The second-order terms and remainders satisfy

$$\mathbb{E}_{\mu_0}\|\widetilde{Q}_0(X)\|^2 + \mathbb{E}_{\mu_1}\|\widetilde{Q}_1(Y)\|^2 \le C_Q,$$

and

$$\|\mathcal{R}_0^\alpha\|_{L^2(\mu_0)} + \|\mathcal{R}_1^\alpha\|_{L^2(\mu_1)} \le C_R\alpha^3.$$

The quantities $\widetilde{W}_0, \widetilde{W}_1$ measure whether the constrained forward and backward action directions are tangent to the inverse-consistency relations $g \circ f = \mathrm{Id}$ and $f \circ g = \mathrm{Id}$. Define the projected mismatch energy by

$$\widetilde{E}_{\mathrm{mismatch}}(z) := \mathbb{E}_{\mu_0}\|\widetilde{W}_0(X)\|^2 + \mathbb{E}_{\mu_1}\|\widetilde{W}_1(Y)\|^2.$$

We define the cycle-stationary transport set by

$$\mathcal{S}_{\mathrm{stat}} := \left\{ z = (f,g) \in \mathcal{M}_{\mathrm{match}} : \mathcal{C}(z) = 0,\ \widetilde{E}_{\mathrm{mismatch}}(z) = 0 \right\}.$$

Thus, $\mathcal{S}_{\mathrm{stat}}$ consists of marginal-matching pairs that are either at the critical point with respect to the transport-loss or the two independent updates satisfy inverse coordination for $\widetilde{W}_0(X) = 0 = \widetilde{W}_1(X)$. The true OT pair belongs to this set. Indeed, since

$$(T^*, S^*) = \operatorname*{argmin}_{(f,g)\in\mathcal{M}_{\mathrm{match}}} \mathcal{A}(f,g),$$

the constrained transport gradient vanishes:

$$\widetilde{U}_f = 0, \qquad \widetilde{U}_g = 0.$$

Moreover,

$$S^* \circ T^* = \mathrm{Id} \quad \mu_0\text{-a.e.}, \qquad T^* \circ S^* = \mathrm{Id} \quad \mu_1\text{-a.e.}$$

Therefore

$$(T^*, S^*) \in \mathcal{S}_{\mathrm{stat}}.$$

LEMMA 3.5.2.1 (Persistent low consistency implies small projected mismatch). *Let $z = (f, g) \in \mathcal{M}_{\mathrm{match}}$ satisfy projected Taylor regularity with constants (α_0, C_Q, C_R). Assume that, for some $\delta, \varepsilon \geq 0$,*

$$\mathcal{C}(z) \leq \delta^2, \qquad \mathcal{C}(z^+) \leq \delta^2 + \varepsilon.$$

Then, for every $0 < \alpha \leq \alpha_0$,

$$\widetilde{E}_{\mathrm{mismatch}}(z) \leq C \left(\frac{\delta^2 + \varepsilon}{\alpha^2} + \delta + \alpha \right),$$

where $C > 0$ depends only on C_Q and C_R.

THEOREM 3.5.3 (Persistent low consistency implies stationarity in residual energy). *Let $z_n = (f_n, g_n) \in \mathcal{M}_{\mathrm{match}}$ satisfy projected Taylor regularity with uniform constants. Suppose that, for some $\delta, \varepsilon \geq 0$,*

$$\limsup_{n\to\infty} \mathcal{C}(z_n) \leq \delta^2, \qquad \limsup_{n\to\infty} \mathcal{C}(z_n^+) \leq \delta^2 + \varepsilon.$$

Let

$$\alpha := (\delta^2 + \varepsilon)^{1/3}$$

satisfying $0 < \alpha \leq \alpha_0$. Then

$$\limsup_{n\to\infty} \widetilde{E}_{\mathrm{mismatch}}(z_n) \leq C \left((\delta^2 + \varepsilon)^{1/3} + \delta \right).$$

PROOF. For every $\eta > 0$, for all sufficiently large n,

$$\mathcal{C}(z_n) \leq \delta^2 + \eta, \qquad \mathcal{C}(z_n^+) \leq \delta^2 + \varepsilon + \eta.$$

Applying Lemma 3.5.2.1 with

$$\widetilde{\delta}_\eta^2 = \delta^2 + \eta, \qquad \widetilde{\varepsilon}_\eta = \varepsilon + \eta,$$

and then taking $\limsup_{n\to\infty}$ and $\eta \downarrow 0$ gives the stated bound. □

Finally, we clarify how $\mathcal{S}_{\mathrm{stat}}$ relates to ordinary constrained critical points of the transport-loss update. In particular, a pair in $\mathcal{S}_{\mathrm{stat}}$ has either small constrained transport loss gradients, such as at local minima or saddle points, or its forward and backward projected action gradients are

coordinated as an infinitesimal inverse-preserving deformation:

$$\widetilde{U}_g(f(x)) = -Dg(f(x))\widetilde{U}_f(x) \qquad \mu_0\text{-a.e..}$$

For independently parameterized forward and backward networks, such exact coordination is unlikely to happen. Therefore, unless this cancellation mechanism occurs, membership in $\mathcal{S}_{\text{stat}}$ corresponds locally to low constrained action gradient. In this sense, non-OT inverse-compatible maps are dynamically fragile: they can remain cycle-stable under independent constrained action updates only if their forward and backward corrections are precisely aligned.

3.5.3. Tractable Guarantee via Cyclic Monotonicity Monitoring. The stationary-set result provides a dynamical stability certificate: persistent low consistency drives the learned pair toward stable inverse transports in residual energy. However, global optimality for the quadratic OT problem cannot be certified solely from this stability condition. In particular, direct certification through the L^2 transport action would require knowing the unknown optimal transport cost. To obtain a practically checkable optimality certificate, we instead use the classical cyclic-monotonicity characterization of quadratic-cost optimal transport.

DEFINITION 3.5.4 (Cyclic monotonicity). *A map $f : \mathcal{X} \to \mathbb{R}^d$ is called cyclically monotone on $\mathcal{X}$ if, for every finite collection $\{x_1, \dots, x_K\} \subset \mathcal{X}$ and every permutation $\sigma \in S_K$,*

$$\sum_{i=1}^{K} \langle f(x_i), x_i \rangle \geq \sum_{i=1}^{K} \langle f(x_i), x_{\sigma(i)} \rangle.$$

Equivalently,

$$\sum_{i=1}^{K} \langle f(x_i), x_{\sigma(i)} - x_i \rangle \leq 0 \qquad \text{for all } K \geq 2,\ \sigma \in S_K.$$

Input Convex Neural Networks (ICNNs) (Makkuva et al., 2020) enforce cyclic monotonicity by parameterizing the transport map as the gradient of a convex potential through a restricted architecture (Amos et al., 2017). In comparison, we keep the transport map unconstrained and monitor cyclic monotonicity on independent validation batches. The resulting monitor estimates a space-average cyclic-monotonicity violation. A small average violation yields a practical certificate that the frozen learned map is close to satisfying cyclic monotonicity on fresh random batches.

In the zero-gap case, cyclic monotonicity characterizes the empirical quadratic OT map under marginal matching. See Appendix B.1.3 for the detailed structural verification.

Given an independent mini-batch

$$B = \{x_i\}_{i=1}^{K} \sim \mu_0^K,$$

define the batch cyclic-monotonicity violation of a fixed learned pair $z = (f_\theta, g_\phi)$ by

$$M_z(B) := \frac{1}{K} \max_{\sigma \in S_K} \sum_{i=1}^{K} \left\langle f_\theta(x_i), x_{\sigma(i)} - x_i \right\rangle.$$

Since the identity permutation gives value 0, we have $M_z(B) \geq 0$. Moreover, $M_z(B) = 0$ if and only if the cyclic-monotonicity inequalities hold on the sampled batch B. We define

$$\widehat{\mathcal{L}}(z; B) := M_z(B).$$

For a fixed map configuration z, define its expected batch cyclic profile of size K by

$$\mathcal{L}_K(z) := \mathbb{E}_{B \sim \mu_0^K} \left[\widehat{\mathcal{L}}(z; B) \right].$$

Given independent validation batches $B_1, \ldots, B_N \sim \mu_0^K$, define the empirical average

$$\overline{\mathcal{L}}_{N,K}(z) := \frac{1}{N} \sum_{n=1}^{N} \widehat{\mathcal{L}}(z; B_n).$$

THEOREM 3.5.5 (Space averaging for a fixed learned pair). *Fix a learned pair $z = (f_\theta, g_\phi)$. Let $B_n \sim \mu_0^K$, $n = 1, \ldots, N$, be independent validation batches. Then*

$$\mathbb{E}\Big[\widehat{\mathcal{L}}(z; B_n) \Big] = \mathcal{L}_K(z).$$

Moreover,

$$\overline{\mathcal{L}}_{N,K}(z) \longrightarrow \mathcal{L}_K(z) \qquad \text{almost surely as } N \to \infty.$$

Consequently, if the frozen model satisfies

$$\limsup_{N \to \infty} \overline{\mathcal{L}}_{N,K}(z) \leq \varepsilon,$$

then

$$\mathcal{L}_K(z) \leq \varepsilon.$$

PROOF. For fixed z, the random variables

$$\widehat{\mathcal{L}}(z; B_1), \ldots, \widehat{\mathcal{L}}(z; B_N)$$

are independent and identically distributed with mean $\mathcal{L}_K(z)$. The unbiasedness statement follows directly from the definition of $\mathcal{L}_K(z)$. The almost-sure convergence follows from the strong law of large numbers, provided $\mathbb{E}|\widehat{\mathcal{L}}(z; B)| < \infty$. The final claim follows by taking the lim sup. □

The result shows that $\mathcal{L}_K(z)$ is upper bounded by persistent random batch estimation of the cyclic monotonicity violation, under K permutation restriction.

3.5.4. Cycle Consistency Suppresses Map-level Collapse. The following observation formalizes the sense in which cycle consistency prevents a learned map from collapsing distinct input points to the same mapped point. For $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$, define the source-side and target-side map-level collapse by

$$\mathcal{C}^0_{\text{map}}(f_\theta) := \mathbb{E}[\text{Var}(X_0 \mid f_\theta(X_0))]\,, \qquad \mathcal{C}^1_{\text{map}}(g_\phi) := \mathbb{E}[\text{Var}(X_1 \mid g_\phi(X_1))]\,,$$

where, for a vector-valued random variable X and Y,

$$\text{Var}(X \mid Y) := \mathbb{E}\big[\|X - \mathbb{E}[X \mid Y]\|^2 \,\big|\, Y\big]\,.$$

Thus $\mathcal{C}^0_{\text{map}}(f_\theta)$ measures the residual uncertainty in X_0 after observing $f_\theta(X_0)$. Denote the two cycle-consistency losses as

$$\mathcal{L}^0_{\text{cyc}}(\theta, \phi) := \mathbb{E}_{X_0 \sim \mu_0}\left[\|g_\phi(f_\theta(X_0)) - X_0\|^2\right],$$

and

$$\mathcal{L}^1_{\text{cyc}}(\theta, \phi) := \mathbb{E}_{X_1 \sim \mu_1}\left[\|f_\theta(g_\phi(X_1)) - X_1\|^2\right].$$

OBSERVATION 3.5.6 (Cycle consistency upper bounds map-level collapse). *Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and let $X_0 \sim \mu_0$, $X_1 \sim \mu_1$. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$, $g_\phi : \mathbb{R}^d \to \mathbb{R}^d$ be measurable maps, and assume that*

all quantities below are finite. Then

$$\mathcal{C}^0_{\mathrm{map}}(f_\theta) \le \mathcal{L}^0_{\mathrm{cyc}}(\theta, \phi),$$

and

$$\mathcal{C}^1_{\mathrm{map}}(g_\phi) \le \mathcal{L}^1_{\mathrm{cyc}}(\theta, \phi).$$

Consequently, for the total cycle-consistency loss

$$\mathcal{L}_{\mathrm{cyc}}(\theta, \phi) := \mathcal{L}^0_{\mathrm{cyc}}(\theta, \phi) + \mathcal{L}^1_{\mathrm{cyc}}(\theta, \phi),$$

we have

$$\mathcal{C}^0_{\mathrm{map}}(f_\theta) + \mathcal{C}^1_{\mathrm{map}}(g_\phi) \le \mathcal{L}_{\mathrm{cyc}}(\theta, \phi).$$

PROOF. We prove the source-side bound; the target-side bound is identical. Since $g_\phi(f_\theta(X_0))$ is measurable with respect to $\sigma(f_\theta(X_0))$, the conditional expectation $\mathbb{E}[X_0 \mid f_\theta(X_0)]$ is the best $\sigma(f_\theta(X_0))$-measurable predictor of X_0 in L^2. Therefore,

$$\mathbb{E}\big[\|X_0 - \mathbb{E}[X_0 \mid f_\theta(X_0)]\|^2\big] \le \mathbb{E}\big[\|X_0 - g_\phi(f_\theta(X_0))\|^2\big].$$

The left-hand side is precisely $\mathcal{C}^0_{\mathrm{map}}(f_\theta)$, while the right-hand side is $\mathcal{L}^0_{\mathrm{cyc}}(\theta, \phi)$. Hence

$$\mathcal{C}^0_{\mathrm{map}}(f_\theta) \le \mathcal{L}^0_{\mathrm{cyc}}(\theta, \phi).$$

Applying the same argument with X_1, g_ϕ, and f_θ gives

$$\mathcal{C}^1_{\mathrm{map}}(g_\phi) \le \mathcal{L}^1_{\mathrm{cyc}}(\theta, \phi).$$

Adding the two inequalities yields the final claim. □

3.6. Experiments

We evaluate the proposed method on five different datasets, spanning synthetic geometry, handwritten images, natural images, and scientific data: *Swiss roll*, *MNIST*, *CelebA*, *single-cell data*, and *chest X-ray images*. The experiments are designed to test three aspects of the method that are central to our claims: (i) *marginal matching*, namely whether the learned pushforward

matches the target distribution; (ii) *transport structure*, namely whether the learned map remains short, regular, and approximately invertible; and (iii) *intermediate geometry*, namely whether the induced interpolation path is more coherent than that of standard baselines. As we show in our results, a consistent pattern emerges across datasets: minimizing transport cost alone, without strong target-marginal enforcement, tends to produce short but distributionally inadequate maps, thus motivating the need for the boundary/intermediate matching terms in our objective.

We compare against five representative baselines: entropic OT (Sinkhorn), neural OT, ICNN-based OT, rectified flow, and mini-batch OT-guided flow matching. Sinkhorn provides a classical coupling-based benchmark with strong finite-sample marginal alignment, but does not directly define a global map for unseen samples. Neural OT provides a learned Monge-map baseline based on a saddle-point objective. ICNN-based OT imposes the Brenier structure by learning convex Kantorovich potentials. Rectified flow and mini-batch OT-guided flow matching provide flow-based transport baselines, with the latter using mini-batch OT couplings to guide endpoint pairings.

3.6.1. Comparison methods. We compare the following methods:

- **Ours**: a bidirectional learned transport method with forward and backward maps, dynamic path matching, cycle consistency, and an annealed quadratic displacement penalty.
- **Entropic OT**: a finite-sample Sinkhorn solver. For image experiments, we use the entropic coupling to obtain hard matches by selecting the target point with largest transport mass for each source point.
- **Neural OT**: a learned deterministic map trained through a neural optimal transport saddle objective.
- **ICNN OT**: an input-convex neural network baseline that learns convex Kantorovich potentials and defines the transport map as a potential gradient.
- **Rectified flow with reflow**: an iterative rectified-flow baseline that repeatedly retrains the velocity field using endpoint pairs induced by the previously learned flow, thereby straightening the learned transport trajectories.
- **Mini-batch OT-guided flow matching**: an OT-CFM-style baseline that first computes mini-batch Sinkhorn pairings and then trains a velocity field on the induced straight-line paths.

3.6.2. Comparison metrics. We use the following metrics for comparing methods. Note that every metric is not used for each experiment. The metrics used for each evaluation are noted in their respective sections.

- Transport cost, as defined in Equation 3.3.1, is the average per-sample squared L_2 distance in pixel space, computed as the sum of the squared pixel-wise differences between each source image and its transported image, averaged over all samples.

$$\text{Transport Cost} = \frac{1}{N}\sum_{i=1}^{N}\|T(x_i) - x_i\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(T(x_i)_{c,h,w} - x_{i,c,h,w}\right)^2.$$

  - N denotes the total number of images used for evaluation;
  - C denotes the number of image channels;
  - H denotes the image height;
  - W denotes the image width.
- Target Domain Rate (TDR) measures the percentage of transported samples that are classified as belonging to the target domain, with a higher TDR indicating more successful domain transfer.
- Target Probability Score (TPS) measures the average probability (confidence) that the classifier assigns to the target-domain classes. A higher TPS indicates that the classifier is more confident in assigning the transported samples to the target domain.
- Fréchet Inception Distance (FID) measures the discrepancy between the distribution of transported samples and that of real target-domain samples in a pretrained feature space. The distance is computed by:

$$\text{FID} = \|\mu_{\text{map}} - \mu_{\text{target}}\|_2^2 + \text{Tr}\left(\Sigma_{\text{map}} + \Sigma_{\text{target}} - 2\left(\Sigma_{\text{map}}\Sigma_{\text{target}}\right)^{1/2}\right).$$

  where μ and Σ denote the empirical mean and covariance matrix, respectively, of the features extracted from the samples. We use `pytorch-fid`'s implementation to calculate FID (Seitzer, 2020).
- Class-distribution Total Variation Distance (Class-TV) measures how closely the class distribution of the generated target samples matches the class distribution of the real

target dataset. For K total classes, let p_k denote the proportion of generated samples predicted as class k, conditioned on the prediction belonging to the target class and let q_k denote the proportion of class k in the real target test set. The metric is defined as

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{k \in K} |p_k - q_k|.$$

A smaller value indicates better agreement between the generated and real target-class distributions.

- Maximum Mean Discrepancy (MMD) measures the discrepancy between the transported distribution and the target distribution. It is computed by Gaussian RBF kernels. $\{x_i\}_{i=1}^n$ denote the target samples and $\{y_j\}_{j=1}^m$ denote the transported samples. For a kernel

$$k_\gamma(u, v) = \exp\bigl(-\gamma \|u - v\|_2^2\bigr),$$

the empirical squared MMD is

$$\widehat{\mathrm{MMD}}_\gamma^2 = \frac{1}{n^2} \sum_{i,i'} k_\gamma(x_i, x_{i'}) + \frac{1}{m^2} \sum_{j,j'} k_\gamma(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k_\gamma(x_i, y_j).$$

We evaluate this quantity over 50 logarithmically spaced kernel parameters $\gamma \in [10^{-3}, 10]$ and report their average. A smaller MMD value indicates closer agreement between the transported and target distributions.

- Enrichment-k100 assesses the mixing between the transported and real target samples. For each transported sample x_i, we identify its k nearest neighbors $\mathcal{N}_k(x_i)$, in the combined transported and target sample set using Euclidean distance, excluding the query sample itself. The enrichment score is defined as

$$\mathrm{Enrichment}_k = \frac{1}{nk} \sum_{i=1}^{n} \sum_{z \in \mathcal{N}_k(x_i)} \mathbf{1}\{z \text{ is transported}\}.$$

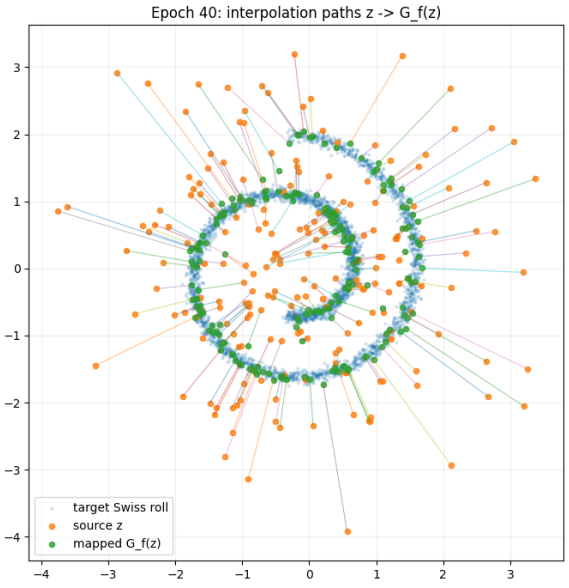
We use $k = 100$. When the transported and target sets have equal sizes, the random-mixing baseline is approximately 0.5. Therefore, values closer to 0.5 indicate better local mixing, whereas larger values indicate that transported samples cluster separately from the real target samples.

- Training time is measured as the elapsed training time required to reach the checkpoint used for quantitative evaluation. It includes the training loop, data loading, optimizer updates, logging, and checkpointing, but excludes post-training evaluation. All methods are trained using the same batch size, learning rate, and optimizer.
- Image Exposure (IE) count how many times real images from the training set are used during training. It provides a standardized measure of sample usage across methods, with fewer exposures indicating greater sample efficiency.

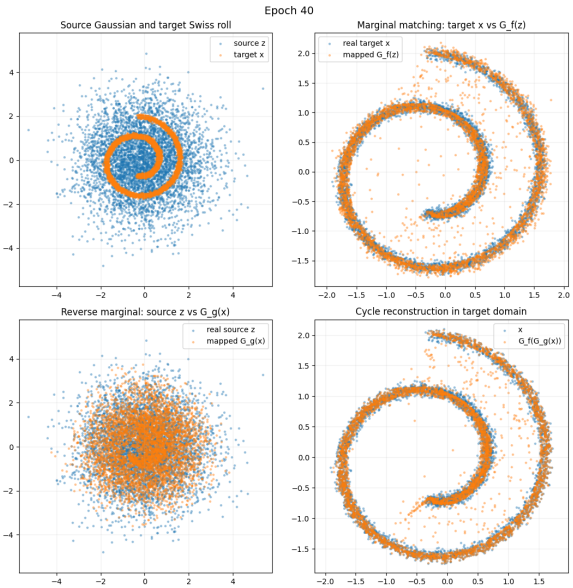
3.6.3. Swiss roll: geometric fidelity in low dimensions. We first consider transport from a two-dimensional Gaussian source to a Swiss-roll target. This experiment isolates the geometric aspect of the problem: both endpoint distributions and the induced transport paths can be visualized directly in the data space. The goal is not only to match the target marginal, but also to learn a transport map whose pointwise displacements are coherent with the geometry of the target manifold.

3.6.3.1. *Results.* Figure 3.2 provides a qualitative visualization of the learned bidirectional transport. The marginal-matching panels show that the forward map transports Gaussian samples toward the Swiss-roll target distribution, while the backward map maps target samples back toward the Gaussian source distribution. The cycle reconstruction panel shows that applying the backward and forward maps in sequence approximately preserves target samples, suggesting that the learned maps are not merely matching marginals but also retain approximate inverse structure. The interpolation-path panel further visualizes pointwise paths from source samples z to their transported endpoints $f_\theta(z)$, revealing the displacement field induced by the learned forward map.

3.6.4. MNIST: transport from digits 0–4 to digits 5–9. We next study unpaired transport between two handwritten digit domains. The source distribution consists of MNIST digits $0, 1, 2, 3, 4$, and the target distribution consists of digits $5, 6, 7, 8, 9$. This experiment tests whether a method can learn a global map that transports samples into the target digit domain while maintaining controlled displacement and avoiding trivial low-cost solutions that remain close to the source images.



(A) Pointwise interpolation paths $z \to f_\theta(z)$.



(B) Marginal matching and cycle reconstruction.

FIGURE 3.2. Qualitative visualization of the Gaussian-to-Swiss-roll experiment. The left panel shows pointwise map-induced interpolation paths from source samples to transported endpoints. The right panel shows forward marginal matching, reverse marginal matching, and target-domain cycle reconstruction.

3.6.4.1. *Setup.* All neural methods use MNIST images normalized to $[-1, 1]$. The proposed method learns a forward map f_θ from digits 0–4 to digits 5–9 and a backward map g_ϕ from digits 5–9 to digits 0–4, both implemented as residual CNN encoder–decoder maps. The dynamic matching term uses the intermediate-time discriminator $D^t_{\omega_t}$, which compares forward and backward interpolation paths.

$$X_t^F = (1-t)x_0 + t f_\theta(x_0), \qquad X_t^B = ty + (1-t)g_\phi(y).$$

The training objective combines this dynamic adversarial matching loss with symmetric cycle consistency and an annealed quadratic displacement penalty. The displacement weight is gradually annealed throughout training, allowing strong geometric regularization early on while progressively relaxing the constraint to preserve final image quality.

For entropic OT, we solve a finite-sample Sinkhorn problem between sampled source images and sampled target images using squared Euclidean pixel cost. Given the entropic coupling P, we

form a hard matching by assigning each source image x_i to the target image $y_{j^*(i)}$, where

$$j^*(i) = \arg\max_j P_{ij}.$$

This baseline is transductive: it provides a strong finite-sample OT reference but does not learn a global map for unseen source images. For neural OT, ICNN OT, rectified flow, and mini-batch OT-guided flow matching, we train global maps or ODE flows and evaluate them on held-out source test images.

3.6.4.2. *Evaluation.* Across experiments, we report metrics that separately quantify transport structure, target-population matching, and computational efficiency. For the MNIST transport task, these include the per-sample squared transport cost, TDR, TPS, MNIST-feature FID, Class-TV, and training time.

3.6.4.3. *Results.* Table 3.1 reports the quantitative results. Among learned global maps, the proposed method achieves the best overall tradeoff between target-domain generation quality and transport structure. It attains a target-domain rate of 0.957, indicating that most transported images are classified as digits 5–9, while also achieving the lowest MNIST-feature FID among learned methods. The class-distribution TV distance is also the smallest among learned methods, suggesting that the transported samples better cover the target digit classes rather than collapsing to a small subset of target digits.

Neural OT achieves a slightly higher target-domain rate, but its MNIST-feature FID and class-distribution discrepancy are substantially worse, indicating weaker target-distribution matching despite high target-domain validity. ICNN OT, rectified flow, and mini-batch OT-guided flow matching achieve lower transport costs, but their target-domain rates are much lower. This indicates that low displacement alone is insufficient: these methods often remain too close to the source domain and therefore fail to reliably transport digits 0–4 into the target digit domain. In contrast, the proposed method accepts a moderate increase in transport cost in exchange for substantially better target-domain validity and target-distribution matching.

The Sinkhorn hard-matching baseline achieves a very high target-domain rate, as expected, because each matched output is an actual target image from digits 5–9. However, this method is finite-sample and transductive rather than a learned global map. Its transport cost is therefore

Method	Global map?	Cost ↓	TDR ↑	TPS ↑	FID ↓	Class-TV ↓	Train Time (min) ↓
Sinkhorn	No	245.201	0.992	0.984	93.773	0.072	–
Ours	Yes	212.43	**0.957**	**0.954**	**92.269**	**0.039**	**7**
Neural OT	Yes	**133.68**	0.956	0.952	120.45	0.130	198
ICNN OT	Yes	210.82	0.805	0.749	1565.65	0.162	43
Rectified flow	Yes	246.18	0.565	0.567	1282.97	0.171	29
OT-CFM	Yes	100.87	0.390	0.393	841.36	0.241	32

TABLE 3.1. MNIST 0–4 → 5–9 transport. We report transport cost, target-domain rate (TDR), target probability score (TPS), MNIST-feature FID, and total variation distance between predicted and real target-class distributions over digits 5–9. Sinkhorn hard matching is a finite-sample transductive baseline and does not define a learned global map. Among learned global maps, the proposed method achieves the best MNIST-feature FID and target class-distribution matching, while maintaining high target-domain validity.



FIGURE 3.3. MNIST 0–4 → 5–9 transport. Each method is evaluated on the same held-out source digits. We show source images and transported outputs for the learned global maps, together with the finite-sample Sinkhorn hard-matching reference.

not directly comparable to the learned-map costs unless computed under the same normalization and averaging convention. We include it as a classical OT reference rather than as a deployable image-to-image map.

Figure 3.3 shows representative transported samples. The visual results are consistent with the quantitative metrics: the proposed method produces samples that are visually closer to target digits 5–9, while the lower-cost flow baselines tend to preserve too much of the original 0–4 digit structure.

3.6.5. CelebA: bidirectional transport on natural images. We then consider an unpaired domain transport task on CelebA, mapping between female and male face distributions.

Compared with MNIST, this setting is substantially more challenging: the data lie on a richer natural-image manifold, and successful transport must preserve shared facial structure while modifying domain-level attributes.

3.6.5.1. *Setup.* We partition CelebA using the `Male` attribute, taking female faces as the source domain and male faces as the target domain. The training objective follows the same bidirectional principle used in the MNIST experiment: adversarial path matching, cycle consistency, and an annealed quadratic displacement penalty.

3.6.5.2. *Evaluation.* Across all experiments, we present qualitative results and quantitatively compare our method with Neural OT using seven evaluation metrics: transport cost, FID, TDR, TPS, Enrich-k100, Female Image Exposures (FIE), and Male Image Exposures (MIE). The results are reported in Table 3.2. Both TDR and TPS are computed using the FairFace classifier (Kärkkäinen et al., 2019). As a real-data reference, we also report FairFace results on real CelebA female and male images in Table 3.3.

3.6.5.3. *Results.* Figure 3.4 qualitatively demonstrates the learned bidirectional transport on CelebA. The forward and backward maps modify gender-related visual attributes while largely preserving pose, facial layout, and background. The generated images do not exhibit obvious visual artifacts, suggesting that the learned maps remain stable under cross-domain translation. The cycle reconstructions $g_\phi(f_\theta(x))$ and $f_\theta(g_\phi(y))$ are visually close to the original samples, indicating that the cycle-consistency term helps retain sample-level information and discourages many-to-one collapse. Although some fine details are blurred or slightly smoothed, the reconstructions preserve the main semantic and structural content of the inputs.

Figure 3.5 shows that, under the same training budget, our method yields stronger and more consistent female-to-male translations than Neural OT. The generated samples from our method exhibit clearer target-domain attributes while preserving the pose, background, and coarse facial structure of the source images. By contrast, Neural OT often produces weaker transformations and retains more source-domain appearance. These qualitative results suggest that our objective improves target-domain alignment while maintaining sample-level transport structure.

Table 3.2 shows that Neural OT achieves a lower transport cost, whereas our method performs better in terms of both TDR and TPS, indicating that the transported female images are more

Method	Cost ↓	FID ↓	TDR ↑	TPS ↑	|Enrich-k100 − 0.5| ↓	FIE ↓	MIE ↓
ICNN OT	2266.7034	305.2227	46.0%	0.4851 ± 0.25	0.3759	3.2M	35.2M
Neural OT	**687.9927**	**17.8291**	79.9%	0.7447 ± 0.27	**0.0228**	17.6M	1.6M
Ours	1069.7815	25.4478	**92.0%**	$\mathbf{0.8602 \pm 0.20}$	0.0357	2.73M	2.73M

TABLE 3.2. Quantitative comparison between Neural OT and our method on the CelebA female-to-male translation task.

Input set	TPS	Predicted male	Predicted female
Real source female	0.1278 ± 0.2304	479 (9.6%)	4521 (90.4%)
Real target male	0.8893 ± 0.2074	4629 (92.6%)	371 (7.4%)

TABLE 3.3. FairFace classifier baseline results on real CelebA source-female and target-male images. Each evaluation set contains 5,000 images. The classifier correctly predicts 90.4% of the source-female images as female and 92.6% of the real target-male images as male.

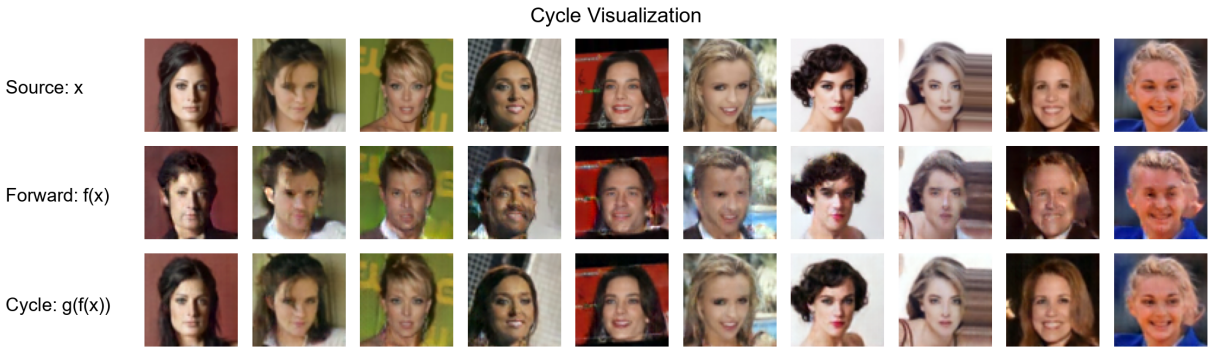
consistently and confidently classified as male. Moreover, our results are closer to the FairFace baseline reported in Table 3.3: our method achieves a TDR of (92.0%), compared with the baseline value of (92.6%), and a TPS of (0.8356 ± 0.2300), compared with the baseline value of (0.8893 ± 0.2074).

3.6.6. Single-cell data: transport on scientific manifolds.

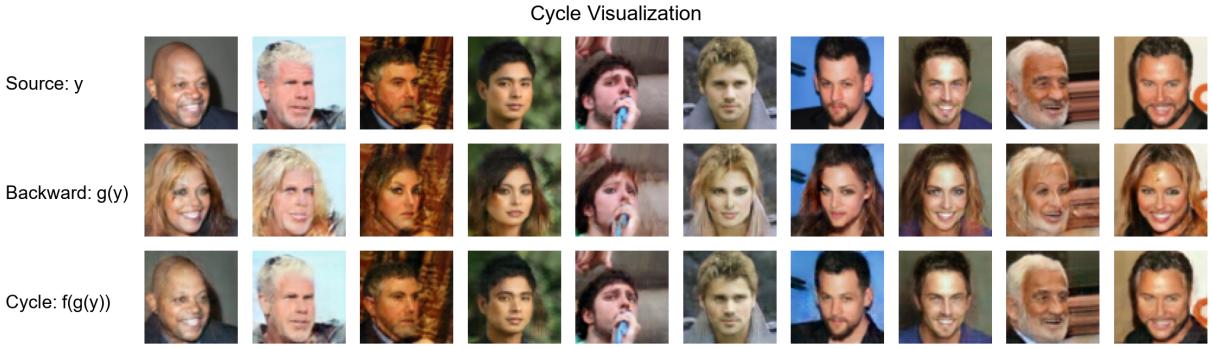
3.6.6.1. *Setup.* We evaluate the methods on a single-cell 4i dataset using the trametinib perturbation experiment. We take cells from the control condition as the source domain and trametinib-treated cells as the target domain. Each cell is represented by the 48 molecular features, and no cell-level correspondence between control and treated cells is assumed. The goal is therefore to learn an unpaired transport map from the control population to the trametinib-treated population. The forward and backward maps are implemented as residual MLPs in the single-cell feature space.

3.6.6.2. *Evaluation.* We evaluate the single-cell transport task using metrics that capture both transport regularity and target-population matching. Specifically, we report transport cost, MMD, and Enrich-k100.

3.6.6.3. *Results.* Table 3.4 summarizes the control-to-trametinib transport results on the 4i dataset. Our method achieves the lowest transport cost among all learned global-map methods,



(A) Female-to-male-to-female cycle.



(B) Male-to-female-to-male cycle.

FIGURE 3.4. Cycle visualizations on CelebA. The first block shows source female samples x, forward mapped samples $f_\theta(x)$, and cycle reconstructions $g_\phi(f_\theta(x))$. The second block shows source male samples y, backward mapped samples $g_\phi(y)$, and cycle reconstructions $f_\theta(g_\phi(y))$.

reducing the average squared displacement to 2.8352, compared with 3.1925 for neural OT and 3.2515 for ICNN OT. This indicates that the learned map reaches the trametinib-treated population with substantially smaller feature-space movement. At the same time, our method remains competitive in target-population matching: its MMD is close to the best value achieved by ICNN OT, and its neighborhood-mixing score and signature correlation are comparable to neural OT and ICNN OT. Neural OT obtains the best local enrichment score and the highest signature correlation, while ICNN OT obtains the lowest MMD; however, both require larger transport costs than our method. In contrast, rectified flow and OT-guided flow matching yield substantially larger transport costs and weaker distributional or perturbation-level matching, suggesting less effective

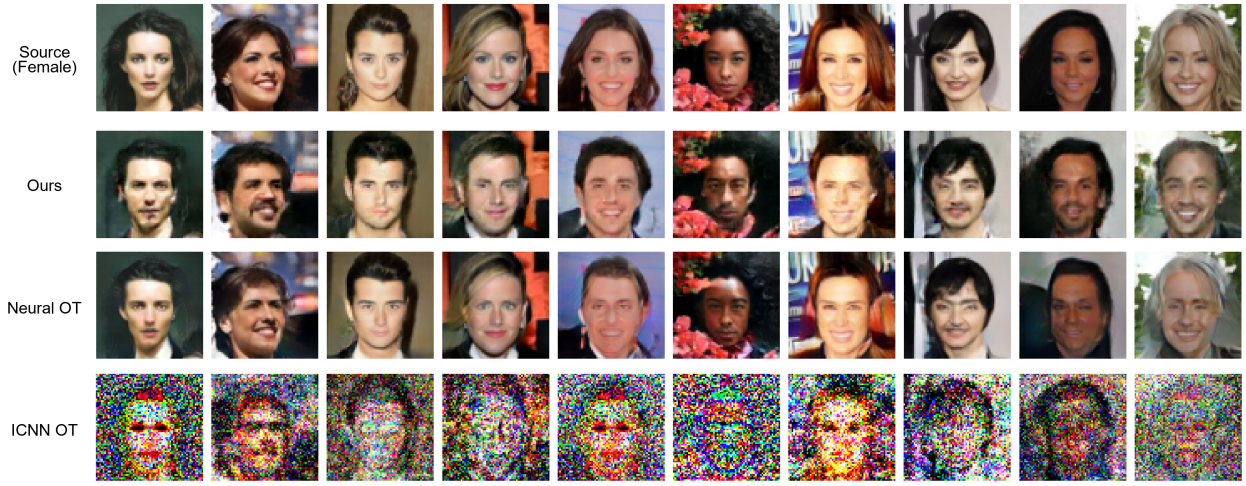


FIGURE 3.5. Comparison results for female-to-male face translation on CelebA.

Method	Cost ↓	MMD ↓	|Enrich-k100 − 0.5| ↓	Signature Corr ↑
Sinkhorn	5.2320	–	–	–
Ours	**2.8352**	**0.002472**	0.00729	**0.99093**
Neural OT	3.1925	0.002879	0.01835	0.98936
ICNN OT	3.2515	0.002623	**0.00456**	0.98698
Rectified flow	9.4423	0.004876	0.02044	0.94781
OT-CFM	7.0150	0.012698	0.11412	0.86522

TABLE 3.4. Results for the control-to-trametini4i trametinib single-cell perturbation transport task, showing transport cost, MMD, enrich-k100, and signature correlation for different transport methods.

transport in this single-cell perturbation setting. Sinkhorn hard matching is included as a finite-sample transductive reference; since it does not define a learned global map for held-out cells, the remaining population-level metrics are not reported.

3.6.7. Chest X-ray Translation: Transport on Medical Images.

3.6.7.1. *Setup.* We further evaluate our method on chest x-ray images to test whether the proposed forward–backward transport framework can operate on even higher-dimensional medical data. Compared with natural image attribute translation, chest radiographs impose a stronger structure-preservation requirement: a meaningful transport map should modify the domain-specific

Method	Cost ↓	|Enrich-k100 − 0.5| ↓	TPS ↑	Training Time (hours:minutes) ↓
Sinkhorn	5643.2637	–	–	**<0:01**
ICNN OT	65095.5671	0.4984	0.1513 ± 0.0409	1:01
Neural OT	1984.1612	0.3246	**0.4762** ± 0.1860	19:15
Ours	**1717.2522**	**0.3169**	0.4098 ± 0.1873	8:41

TABLE 3.5. Quantitative comparison between ICCN OT, Neural OT, and our method on the chest x-ray healthy-to-pneumonia transition task.

Input set	TPS
Real source (healthy)	0.1876 ± 0.1382
Real target (pneumonia)	0.6254 ± 0.1902

TABLE 3.6. Pneumonia classifier baseline results on real source (healthy) and target (pneumonia) images.

radiographic appearance while preserving the underlying anatomical layout, including the lung fields, ribs, clavicles, and global patient pose.

We use data from the RSNA Pneumonia Detection Challenge[1] for training (X. Wang et al., 2017). The dataset includes a total of 7,106 chest x-rays from patients with pneumonia and 9,790 x-rays from patients without pneumonia. We use a random subset of 3,000 images from each class for training, and repeat the training three times for each method. The full dataset is used for evaluation.

3.6.7.2. *Evaluation.* To evaluate the utility of our and other methods at converting x-ray images between the two classes we use a classification model developed by Weng et al. (2017) on the dataset provided by Rajpurkar et al. (2017). We report the transport cost, Enrich-k100, TPS, and total training time of each method. The results in Table 3.5 show the mapping from healthy images to ones with pneumonia. The results for the mapping in the other direction can be found in B.3.1.

3.6.7.3. *Results.* Figure 3.6 shows representative qualitative results. The first row shows source chest X-ray images x, the second row shows transported images $f_\theta(x)$, and the third row shows cycle reconstructions $g_\phi(f_\theta(x))$. The transported samples exhibit visible changes in global radiographic appearance, including contrast, opacity, and local texture patterns, while largely preserving the

[1]Data is available to download here: https://nihcc.app.box.com/v/ChestXray-NIHCC

coarse anatomical geometry of the input images. In particular, the position of the thoracic cavity, lung fields, shoulders, and mediastinal structure remains aligned with the corresponding source image. The cycle reconstructions recover the main image structure and much of the source-domain appearance, suggesting that the learned map is not simply collapsing different inputs to a small set of target-like images.

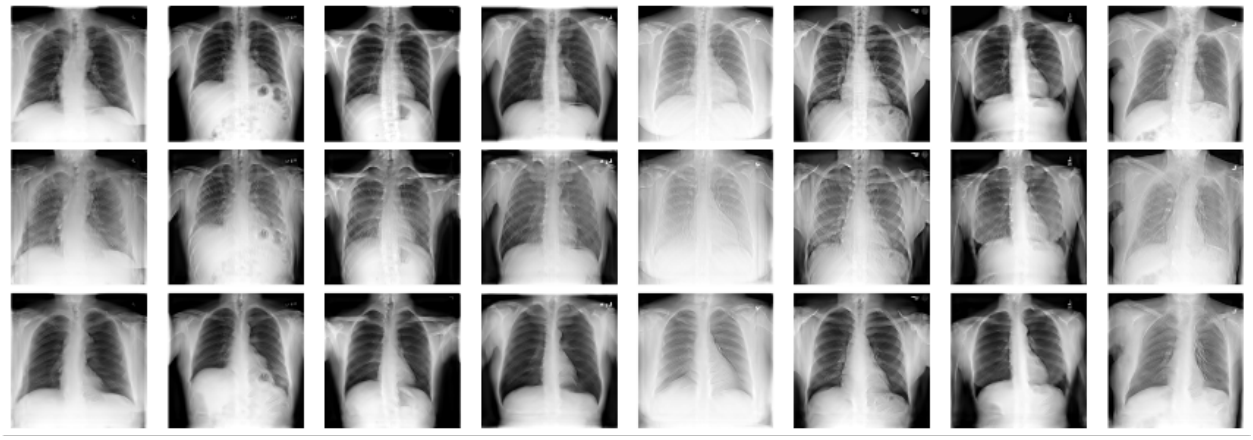
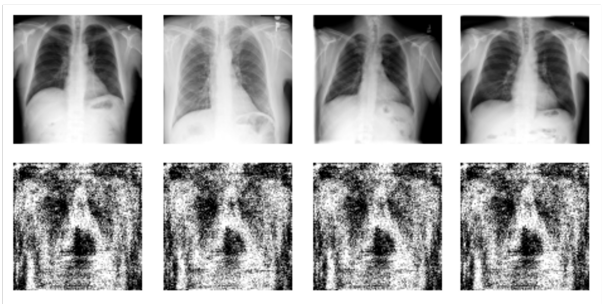
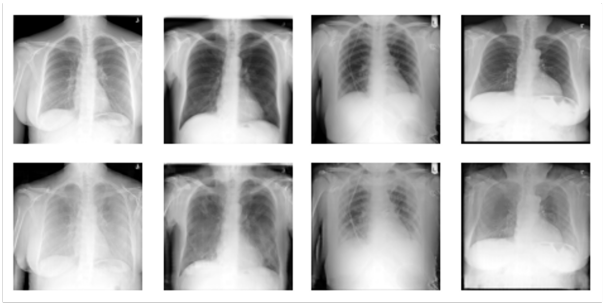


FIGURE 3.6. Chest x-ray transport results for our method. The first row shows source (healthy) radiographs x, the second row shows the transported samples $f_\theta(x)$, and the third row shows the cycle reconstructions $g_\phi(f_\theta(x))$. The learned transport changes the target-domain radiographic appearance while preserving the coarse anatomical layout.



(A) Images created by a model trained with ICNN OT.



(B) Images created by a model trained with Neural OT.

FIGURE 3.7. Chest x-ray transport results for other tested methods. The first row shows source (healthy) radiographs x and the second row shows the transported samples $f_\theta(x)$.

3.6.8. Takeaway. Taken together, the experiments illustrate a consistent pattern. Namely, transport cost by itself is not a reliable proxy for successful transport when the target marginal is not enforced strongly. On MNIST and CelebA, several baselines (ICNN OT, rectified flow, and mini-batch OT-guided flow matching on MNIST; Neural OT on CelebA) achieve lower transport cost than our method but substantially weaker target-domain matching, indicating that these maps remain too close to the source distribution rather than genuinely reaching the target. This motivates the explicit boundary (and, under the revised objective, intermediate) marginal-matching terms in our loss, rather than relying on transport-cost minimization alone. Entropic OT remains a strong classical baseline for finite-sample marginal alignment, especially in low dimension, but it does not directly provide a learned global map and may require additional projection or regression steps. Neural OT provides expressive learned maps, but without explicit reverse consistency it may admit transport solutions that match target-domain labels while failing to match the full target distribution. ICNN OT encodes the Brenier structure but is difficult to optimize directly in high-dimensional pixel space. Flow-matching baselines provide flexible ODE transports, but without sufficiently informative couplings they may favor low-displacement maps that do not fully reach the target domain. The proposed method combines the flexibility of learned transport with bidirectional structural regularization, leading to maps that are not only distributionally accurate but also more stable and coherent along intermediate paths.

3.7. Conclusion

We introduced a bidirectional dynamic optimal transport framework that learns quadratic OT maps from unpaired samples without precomputed couplings, minibatch or semi-discrete OT solvers, or convex-potential parameterizations. The key idea is to construct synchronized forward and backward McCann-type interpolants and to enforce their agreement at intermediate times, turning target-free intermediate supervision into an explicit training signal rather than an artifact of a prescribed coupling. We showed that this construction admits a tractable recovery guarantee (Theorem 3.4.1): for sufficiently small transport weight λ, the global minimizer of our population

objective is arbitrarily close to the true Brenier map and its inverse in L^2. We further gave a complementary dynamical account of why cycle consistency and transport-cost minimization interact productively rather than merely trading off against one another (Section 3.5), and showed empirically, across synthetic geometry, handwritten digits, natural face images, single-cell perturbation data, and chest radiographs, that the resulting maps achieve competitive or superior target-distribution matching at moderate transport cost, without the instability or restrictive architectures associated with existing Brenier-potential or minibatch-coupling approaches.

3.7.1. Limitations. Our theoretical guarantees are stated under idealized conditions – compact domains, absolutely continuous marginals, and sufficiently expressive generator and discriminator classes – and describe the population-level minimizer rather than the finite-sample, finite-capacity solution reached by stochastic optimization. In practice, the adversarial JSD estimator used for boundary and intermediate matching inherits the familiar optimization sensitivities of GAN-style training, and the annealing schedule for the transport weight λ_k is a practical heuristic for approaching the small-λ regime required by Theorem 3.4.1, rather than a schedule whose convergence is itself established by the theorem. Our cyclic-monotonicity monitor (Section 3.5.3) offers a checkable diagnostic rather than a certificate of global optimality on its own. Finally, our current formulation targets the quadratic transport cost between two fixed marginals. Extending the framework to general ground costs, conditional or multi-marginal transport, and settings with side information or paired covariates are natural directions we have not explored here.

3.7.2. Future work. Promising directions include replacing the adversarial intermediate-time and boundary discriminators with non-adversarial discrepancy estimators (e.g., kernel-based MMD or sliced-Wasserstein distances) to avoid the optimization instabilities inherent in minimax training, deriving finite-sample (rather than population-level) recovery guarantees that account for network capacity and optimization error, and extending the bidirectional interpolant construction to non-quadratic costs and multi-marginal or Schrödinger-bridge-type settings. On the empirical side, expanding experiments to include additional data types or domains such as longitudinal or multi-condition single-cell data would further test the practical reach of target-free bidirectional supervision.

CHAPTER 4

The "I Don't Know" Filter: Enhancing Agentic Reliability in Function Calling

This chapter marks a shift in focus from medical AI to large language models. It is based on a paper that will appear at the 2026 IEEE International Conference on Systems, Man, and Cybernetics (and available on arXiv here) and is a collaboration with Mason del Rosario and Boris Selitser from Okareo, Inc. and Thomas Strohmer from the UC Davis Department of Mathematics. This work develops a metric that better captures the stakes of tool calling, and a technique that quantifies uncertainty in LLM-based agentic tool use in order to improve the reliability of agents that act on the world directly. Agentic tool use, where an LLM interacts with tools or APIs rather than a human, is a narrow slice of LLM capability, but an increasingly consequential one: because an agent acts without a human in the loop, and its actions are often invisible to the end user, trusting an agent means trusting that it is doing exactly what the user expects. Minimizing that uncertainty is a key step toward building that trust.

4.1. Abstract

The language models that underpin agents have seen a rapid rise in performance on function calling benchmarks. However, the metrics used in the training and evaluation of these models often encourage models to make positive claims even when the answer is uncertain, leading to hallucinations. Such hallucinations can be disastrous when language models are trusted to use function calls to make decisions in high stakes applications. To that end, we propose an agent evaluation metric that takes into account the negative outcomes associated with incorrect function calls. Further, to catch hallucinations before they can cause harm, we propose a lightweight trainable filter that can quantify a language model's uncertainty and remove potentially harmful function calls. By training that filter to detect and suppress uncertain function calls without modifying the underlying model,

we demonstrate a practical path toward agents that know when to say "I don't know," a property we argue is essential to production reliability.

4.2. Introduction

Agents are language models (LMs) that combine tools, memory, and reasoning to perform tasks (L. Wang et al., 2024). As agents have improved at tool use, developers in critical industries such as digital infrastructure (Hatami et al., 2026), health care, and public services (Deloitte, 2026) have begun to trust agents to make decisions in production settings. While an errant tool call may be a nuisance for a chatbot, such mistakes can have negative externalities when agents are trusted to make critical decisions in production environments (Ray, 2026).

A large scale study of agents in production showed a persistent gap in agent performance on benchmarks and in production systems (Pan et al., 2026). In that study, reliability was most commonly citepd as the bottleneck to agent deployment. A key part of that reliability gap is hallucinations, and given agents' ability to take autonomous action in real-world decision making, hallucinations can cause much more severe and material damage than standalone LMs.

(Kalai et al., 2025) makes the claim that hallucinations persist because of conventions in LM evaluations. In most evaluations, an incorrect answer is weighed the same as a non-answer, so guessing when uncertain improves accuracy. However, Bastounis et al. state that this tendency to provide a non-answer when uncertain is a fundamental roadblock to "trustworthy" (i.e., reliable) AI, and they describe this tendency to hallucinate and the inability to state "I don't know" as a critical component of the Consistent Reasoning Paradox (Bastounis et al., 2024).

In this work, we study "agent reliability" through this lens: the agent's ability to express uncertainty, i.e., by explicitly stating "I don't know" (IDK). Towards this end, we make the following contributions:

- An accuracy metric that penalizes incorrect function calls, emphasizing the danger of an autonomous agent guessing at an answer instead of admitting uncertainty and encouraging IDK responses
- A simple filter that uses "whitebox," "graybox," and "blackbox" features to identify likely-incorrect function calls (cases of IDK) to reduce the likelihood of hallucinations

- A simple method for generating synthetic data that can be used to train that filter, enabling deployment even when labeled function calling data is limited
- A study of the filter's effectiveness in reducing uncertainty across different open source LMs and function calling benchmarks

With these contributions, we demonstrate that explicitly measuring an agent's uncertainty and modulating its output accordingly can reduce the likelihood of incorrect function calls, improving its reliability.

4.3. Background

4.3.1. Uncertainty quantification for LMs. Existing work has applied uncertainty quantification to measure the specific sources of uncertainty in the output of LMs ((Xiaoou Liu et al., 2025; Shorinwa et al., 2025; Lin et al., 2024)). These works have mostly geared towards hallucination detection in freeform, linguistic LM responses rather than function calls specifically. Works in this field use various metrics to quantify the variability of LM outputs. Of note for the present work, Li et al. calculate a so-called "semantic volume," or the volume of the parallelepiped formed by the embeddings of outputs (X. Li et al., 2025). Huang et al. (Huang et al., 2025) use a formulation of variation ratio, adapted by Z. Wang et al. (2022) to be used on LM outputs, to measure the statistical dispersion of outputs.

In work more specifically focused on function calling, Rabanser et al. (2026) break down agent reliability into four dimensions: consistency (repeatable behavior across runs), robustness (stability under input and environmental perturbations), predictability (calibrated confidence and discrimination of correct/incorrect predictions), and safety (bounded severity when failures occur).

While we claim that this work has beneficial implications in all of these domains, our methods focus on **consistency** across different sampling runs. More specifically, we investigate an agent's consistency under "prediction uncertainty", i.e. the uncertainty in an LM's output given an identical input that arises from both variability in the training data (i.e., aleatoric uncertainty) and in genuine gaps in the LM's knowledge (i.e., epistemic uncertainty) (Xiaoou Liu et al., 2025).

4.3.2. Reliable function calling for agents. Existing work on function calling focuses on improving the standard accuracy measures that do not account for an LM's uncertainty. Further,

when dealing with agent reliability, work in this area focuses on **robustness**, or the ability to generate the correct function call under input perturbations (Rabinovich et al., 2025; W. Wang et al., 2023).

A work with more direct relevance to the current work is τ-bench (S. Yao et al., 2024), a benchmark which aims to measure **consistency**. τ-bench uses a metric of agent reliability over repeated trials on the same task which Yao et al. call pass^k. They define pass^k as the chance that all k i.i.d. task trials are successful, averaged across tasks. While pass^k is a valid indicator of LM reliability, the metric does not penalize highly confident yet incorrect answers (i.e., hallucinations), a shortcoming which we address with our proposed metric.

4.3.3. Language model features. To train a filter that measures an LM's uncertainty on a given input text, we need to select a feature set. In this study, we consider features of varying "invasiveness", which we call **whitebox**, **graybox**, or **blackbox**.

In this work, we opt for small language models (SLMs) that can be run locally and allow access to **whitebox** features, or the internal state of the LM (i.e., model parameters, activation values, etc.). In (Marks et al., 2024), Marks et al. show that internal states can reliably encode things like truthfulness, and in (Orgad et al., 2025), Orgad et al. use the internal states of LMs to detect hallucinations.

While access to internal state is an upside of SLMs, it is natural to question their performance relative to large language models (LLMs). However, there is evidence to suggest that SLMs can be just as powerful, if not more powerful, than LLMs when the "objective is schema- and API-constrained accuracy rather than open-ended generation" (Sharma et al., 2025).

While having access to whitebox features is ideal, many production settings only allow the use of **blackbox** features of LMs (i.e., the generated text outputs). Occasionally, such API-gated LMs also provide access to their log probabilities ("logprobs"), and since these logprobs offer a glimpse of the LM's internal state at its outer boundary, we refer the features derived from them as **graybox** features.

In this study, we consider using the full range of whitebox, graybox, and blackbox features in our feature selection for the uncertainty filter, and we analyze the relative importance of each in measuring the LM's uncertainty.

4.3.4. Our Work. Our work builds on two of the key observations mentioned in this section. First, uncertainty quantification research demonstrates that variability in LM outputs can serve as a reliable signal of model uncertainty. Second, existing function calling benchmarks, while measuring accuracy and consistency, do not penalize incorrectness, leaving a gap between benchmark performance and the reliability demands of production agents. We bridge these ideas by treating output inconsistency across repeated function calls as a proxy for agent uncertainty, and by training a classifier to act on that signal by filtering out likely-incorrect calls and replacing them with explicit abstentions. The following section describes this approach in detail.

4.4. Methods

Identifying when an LM agent should abstain from answering a question rather than guessing is central to improving the reliability of agentic systems. To that end, we introduce two concepts: a new evaluation metric and a trained classifier filter. We then validate both across multiple benchmarks and LM agents. The sections below describe the metric, the classifier design, and the experiments used to evaluate them[1].

4.4.1. IDK Score. To better quantify the effectiveness of our classifier in improving the reliability of an agent, we introduce the *I Don't Know Score* (IDKS). IDKS takes into account the negative effect caused by producing an incorrect function call. We define it as:

$$IDKS = \frac{C - I}{N},$$

where C is the number of correct function calls, I is the number of incorrect function calls, and N is the total number of function calls. IDKS ranges from -1 to 1, where -1 means every function call is incorrect and 1 means that every function call is correct. Notably, IDKS does not include agent outputs that indicate uncertainty, like "I don't know." These abstentions are counted as neutral in the calculation of IDKS. Further, we acknowledge that the choice of -1 as the penalty for an incorrect function call in IDKS is a simplification that practitioners can adjust based on their use case.

With this metric in hand, we can now describe the filter designed to improve it.

[1]All code and prompts used in this chapter can be found at https://github.com/SBroecker/idkscore.

4.4.2. Classifier Design. The core of our approach is a classifier trained to distinguish reliable agent outputs from uncertain ones using features derived from repeated generations. At inference time, if the classifier's uncertainty estimate for a given set of outputs exceeds a threshold, the agent's output is filtered out and replaced with an abstention. This concept is illustrated in Figure 4.1.

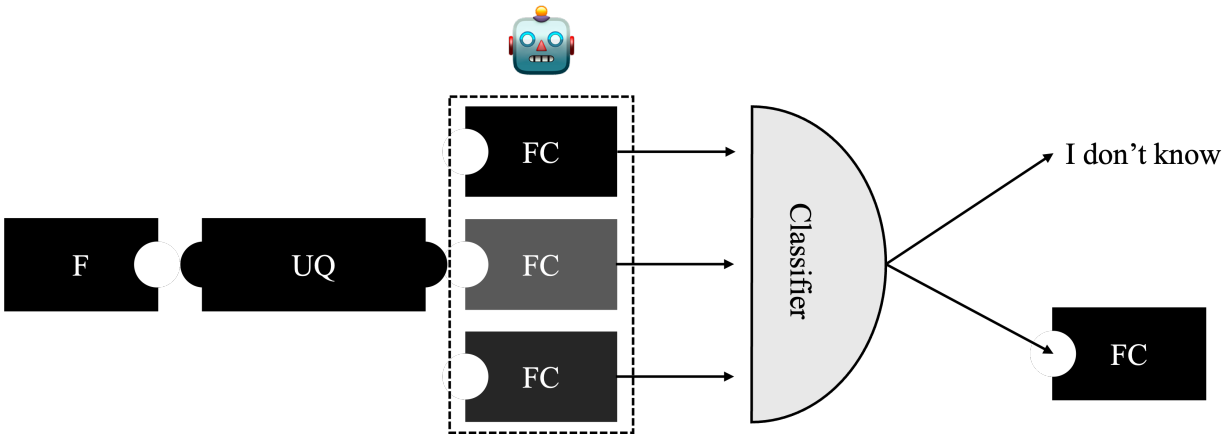


FIGURE 4.1. A classifier being used as a filter for an LM agent. Given the same input functions (F) and user question (UQ), an agent generates multiple function calls (FC). Those repeated function calls are fed into the classifier, which decides whether to use the output or filter it out and respond with "I don't know".

The following sections describe the data and feature extraction pipeline used to train and validate this classifier.

4.4.3. Experimental Setup.

4.4.3.1. *Models.* We evaluate our classifier filter using three open-source SLMs that can be run locally using Huggingface's Transformers library (Wolf et al., 2020), enabling access to whitebox and graybox features: Microsoft's *Phi-4-mini-instruct*, Alibaba Cloud's *Qwen2.5-3B-Instruct*, and Meta's *Llama-3.2-3B-Instruct* (Microsoft et al., 2025; A. Yang et al., 2024; Grattafiori et al., 2024). These models were selected because they contain a similar number of parameters but span a range of model families, allowing us to assess the filter's portability across architecturally distinct agents. Llama and Qwen models are used with default decoding parameters. For Llama this means a temperature of 0.9 and top_p value of 0.6. For Qwen this means a temperature of 0.7, a top_p value of 0.8, and a top_k value of 20. Since Phi models use greedy decoding by default, and since

a greedy decoding strategy doesn't produce any output variability when given the same inputs, we use Llama's default decoding parameters for the Phi model.

4.4.3.2. *Training Data.* We train our classifier on a dataset consisting of synthetic user questions. Our data are created using several Berkeley Function Calling Leaderboard (BFCL) datasets as a starting point (Patil et al., 2023). Specifically, we use BFCL's Simple Python (single functions and function calls), Parallel (single functions with multiple expected function calls), Multiple (multiple possible functions with a single expected function call), and Parallel Multiple (multiple possible functions with multiple expected function calls). While we use BFCL as our starting point, we note that our methodology is generalizable and could thus be used with any existing function calling dataset.

Each datapoint in the BFCL datasets consists of: available functions and their definitions, a user question, and expected function calls. We use that same structure for all of our data. To generate synthetic data, for each datapoint in the aforementioned BFCL datasets, we prompt OpenAI's *gpt-4.1-mini* to create a new user question that would lead to the possible answers provided in the datapoint given the available functions. This is done five times for each datapoint to create five unique variations of user questions.

We call this prompting approach constrained prompting, which we compare to more open-ended prompting, where, for example, an LLM might be prompted to generate both a question and possible function calls given a set of available functions. We found that open-ended prompting produced outputs where the generated questions frequently failed to align with the expected function calls, making them unsuitable as a training data source. An illustrated overview of our constrained prompting approach and an alternative open-ended prompting approach can be seen in Figure 4.2.

Having established the training set, we now describe the held-out data used to evaluate classifier performance.

4.4.3.3. *Validation Data.* We use the aforementioned single-shot BFCL datasets as validation in our experiments: Simple Python, Parallel, Multiple, and Parallel Multiple. We additionally use Simple Java (single functions and function calls with Java formatting), Simple JavaScript (single functions and function calls with JavaScript formatting), and Irrelevance (none of the function choices provided are relevant to the user query and none should be invoked). We also use a subset

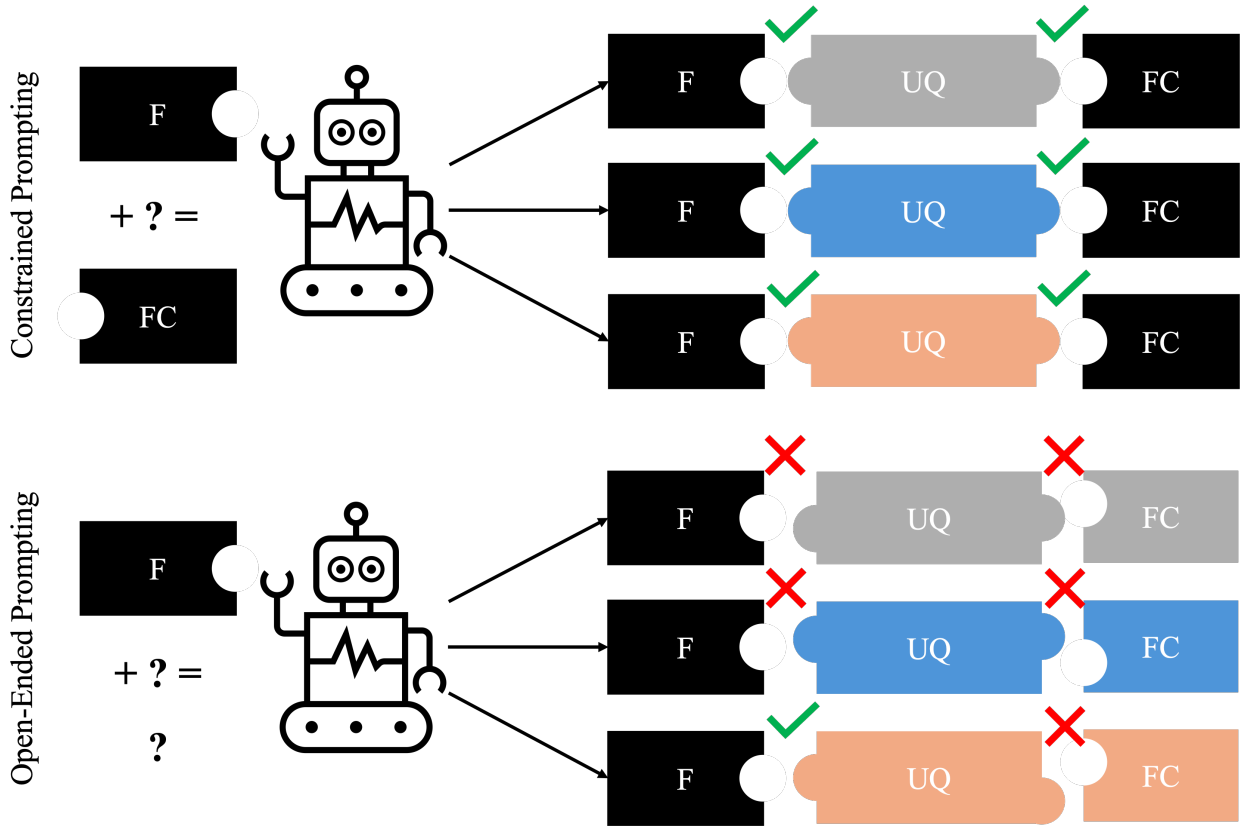


FIGURE 4.2. Synthetic data is generated using constrained prompting, where an LLM is given functions (F) and the expected function calls (FC) and is expected to create valid user questions (UQ). In contrast, open-ended prompting, where the LLM is given less predefined information, produces less reliable function calling scenarios, where the functions, questions, and function call sometimes don't agree with each other.

of datapoints from APIGen, a dataset of synthetically generated function calling examples (Zuxin Liu et al., 2024). To create that subset, we clean it and filter it, and then reformat it to match the format of BFCL datapoints. We then take a random subset of 10,000 examples.

With both datasets defined, we now define the features used to train the classifier.

4.4.3.4. *Feature Extraction.* Features for the classifier are created using the differences observed over repeated generations by the agent. Three different levels of features are extracted. In order of decreasing invasiveness for the LM agent, features are extracted from the final internal state of the agent (whitebox features), the logprobs during generation (graybox features), and the text that

was generated (blackbox features). For blackbox features, downstream features are calculated after the text is embedded using the text embedding model trained by (Y. Liu et al., 2025).

Features are generated from raw whitebox and blackbox features using similar techniques. Namely, semantic volume and variation ratio, two measures of the spread of the features across repeats, are calculated from raw feature vectors (X. Li et al., 2025; Huang et al., 2025). Statistical features are then extracted from the pairwise normalized distances between agent responses. Further statistical features, such as minimum probability and entropy, are generated from raw graybox features.

Labels for every datapoint are created using the function call validation tools provided by (Patil et al., 2023). Since groups of function calls are used for training, we use a variation of the pass^k metric defined by (S. Yao et al., 2024) to label our data, where a group of responses is considered valid if and only if every response in the group is valid.

Before classifier training, irrelevant features are filtered out based on a calculation of the mutual information between each feature and the label. Features are further filtered during training using recursive feature elimination. Both of these filtering steps are performed with tools provided by (Pedregosa et al., 2011). The trained classifier is a random forest model.

The decision threshold for the classifier, above which an agent output is filtered and replaced with an abstention, is selected by maximizing IDKS across benchmarks on the training set. Figure 4.3 shows the relationship between IDKS and the decision threshold, demonstrating the tradeoff between too permissive and too restrictive of a filter. While optimizing on training data risks overfitting the threshold, we deliberately avoid using validation data for this selection to prevent leakage into our held-out results.

The following section describes the three sets of experiments used to characterize the behavior and limitations of our trained classifier.

4.4.4. Experiments. We run three main sets of experiments to verify the effectiveness of our filter design.

4.4.4.1. *Effect of the Number of Repeated Calls.* First, we run experiments varying the number of repeated function calls that are used to build features, ranging from two repeats to five repeats.

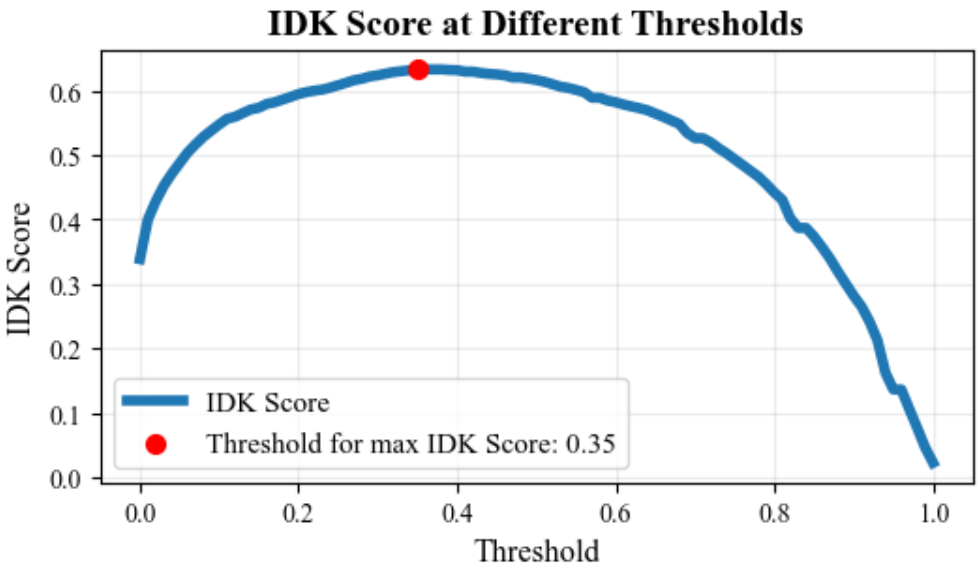


FIGURE 4.3. Average performance of a Llama agent with a classifier filter applied at different decision thresholds. Performance suffers at low thresholds because not enough function calls are filtered and at high thresholds because too many function calls are filtered.

4.4.4.2. *Effect of Feature Invasiveness.* Next, we run experiments varying the level of invasiveness of features. In increasing level of invasiveness, we use: blackbox features only, blackbox and graybox features, and then blackbox, graybox, and whitebox features.

4.4.4.3. *Cross-LM Generalizability.* Finally, we run experiments to test the generalizability of features by training a classifier on the outputs of one LM agent and applying that classifier to the outputs of another LM agent.

4.5. Results

We evaluate our classifier filter across three axes: how performance scales with the number of repeated calls used to build features, how much each additional level of feature invasiveness contributes to model performance, and whether a classifier trained on one LM's outputs transfers to another. In each case, performance is reported using IDKS, where a higher score reflects a more reliable agent.

4.5.1. Effect of the Number of Repeated Calls. Experiments varying the number of repeats show two trends. First, as observed by (S. Yao et al., 2024), agents are more likely to

produce an incorrect function call as you increase the number of function call repeats. This is indicated by a general decrease in baseline IDKS as the number of repeats increases, as can be seen by the white circles in Figure 4.4. Second, the boost in performance created by the trained classifier acting as a filter often increases as you increase the number of repeats, as shown by the solid circles in the figure. However, in a number of the benchmarks, including APIGen and BFCL's Parallel dataset, the boost in performance caused by increasing the number of function call repeats is offset by the degradation of the model's baseline performance, creating relatively constant performance across repeats.

4.5.2. Effect of Feature Invasiveness. Experiments varying the invasiveness of features indicate a boost of performance for each step up in model invasiveness. Table 4.1 shows the results of these experiments. Across most validation datasets, performance increases as you move from blackbox to blackbox and graybox and then to blackbox, graybox, and whitebox features. BFCL's Parallel, Parallel Multiple, and Simple JavaScript, are exceptions to this, where the addition of whitebox features hurts performance slightly. BFCL's Simple Java is another exception, where performance decreases as gray- and whitebox features are added. An apparent outlier, we believe this result is a consequence of the Llama agent's general poor performance on Simple Java and a subsequent lack of a reliable signal.

4.5.3. Cross-LM Generalizability. Experiments across LM agents indicate the portability of the features used to train our classifiers. Figure 4.5 shows the averaged performance across all validation datasets of each LM agent with and without a classifier filter. Further, it shows the agent's performance when the filter is trained on features created by its own LM and different LMs. Performance is strongest for Phi and Qwen agents when the classifier is trained on the LM's own features, with performance slightly dropping when the classifier is trained on a different LM's features. The Llama agent is the exception, where a filter trained on Phi features is the strongest. Further work could explore whether this is a fluke or whether Phi's outputs are a particularly informative signal.

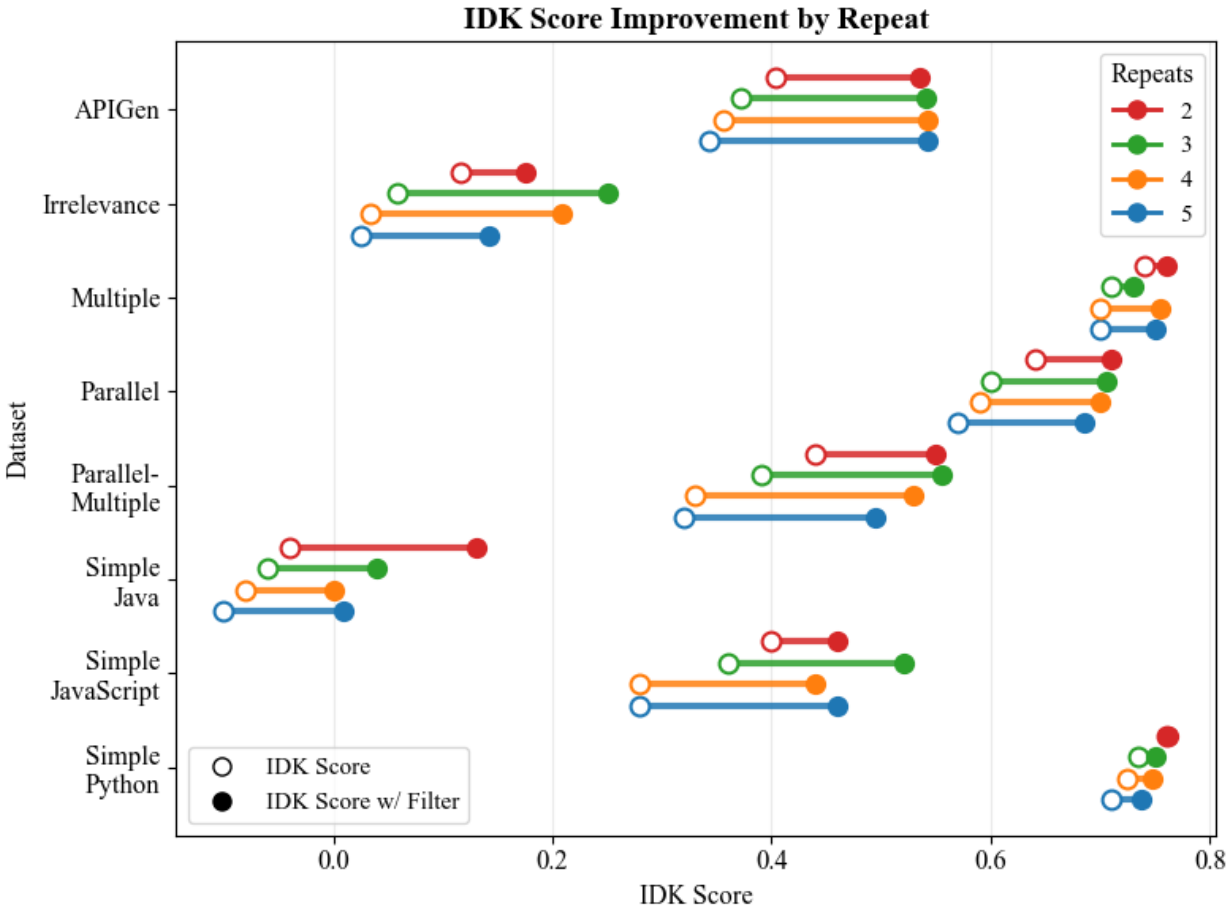


FIGURE 4.4. Llama agent performance across all validation datasets, as measured by IDK Score, with and without a trained classifier acting as a filter. Results are grouped along the y-axis according to validation dataset. Within a group, results are colored by the number of repeated function calls. Each colored line represents the difference in IDK Score between the agent's performance without a classifier filter (white dots) and with a classifier filter (colored dots). Longer lines indicate a larger performance gap.

4.6. Discussion

Our results demonstrate that a consistency-based classifier filter is an effective mechanism for improving the reliability of LM agents in function calling settings. Across benchmarks and model families, applying the filter yields consistent gains in IDKS, confirming that output inconsistency across repeated generations is a useful proxy for agent uncertainty. Crucially, the portability of classifiers across LM agents, where performance drops only slightly when a classifier trained on

	Baseline	Black	Black + Gray	Black + Gray + White
APIGen	0.342	0.453	0.532	0.541
Irrelevance	0.025	0.042	0.217	0.229
Multiple	0.700	0.725	0.750	0.760
Parallel	0.570	0.625	0.680	0.670
Parallel Multiple	0.320	0.465	0.505	0.490
Simple Java	-0.100	0.050	0.000	-0.010
Simple JavaScript	0.280	0.360	0.400	0.380
Simple Python	0.710	0.685	0.738	0.743

TABLE 4.1. IDK Score of a Llama agent trained on Llama outputs with varying levels of feature invasiveness. Performance is shown on different validation datasets. Black, gray, and white refer to blackbox, graybox, and whitebox features, respectively, as described in Section 4.4.3.4.

one model's outputs is applied to another, suggests that the uncertainty signal captured by our features reflects general properties of function calling difficulty rather than idiosyncrasies of any single model.

4.6.1. Consistency Trends Relative to Prior Work. Our results also shed light on how agent consistency has improved since (S. Yao et al., 2024). In that work, repeated sampling led to rapid degradation in pass^k, with model performance dropping by almost half by the fourth repeat for some models. In our experiments, we observe a meaningfully slower degradation rate of $\sim$10% by the fifth repeat for most benchmarks, suggesting that the LMs used in this study have become more consistent in their function calling behavior. However, our results show that even this reduced degradation rate is not negligible: baseline IDKS still declines as the number of repeated calls increases (Figure 4.4), reinforcing that consistency-based filtering remains a valuable intervention even for more capable models.

4.6.2. Failure Modes. Despite its effectiveness, the filter has two notable failure modes that stem from the fundamental limitation of using inconsistency as a proxy for uncertainty.

The first is the case of confident incorrectness: when a model consistently produces the same wrong function call across all repeated generations. In this case, the classifier observes low variance

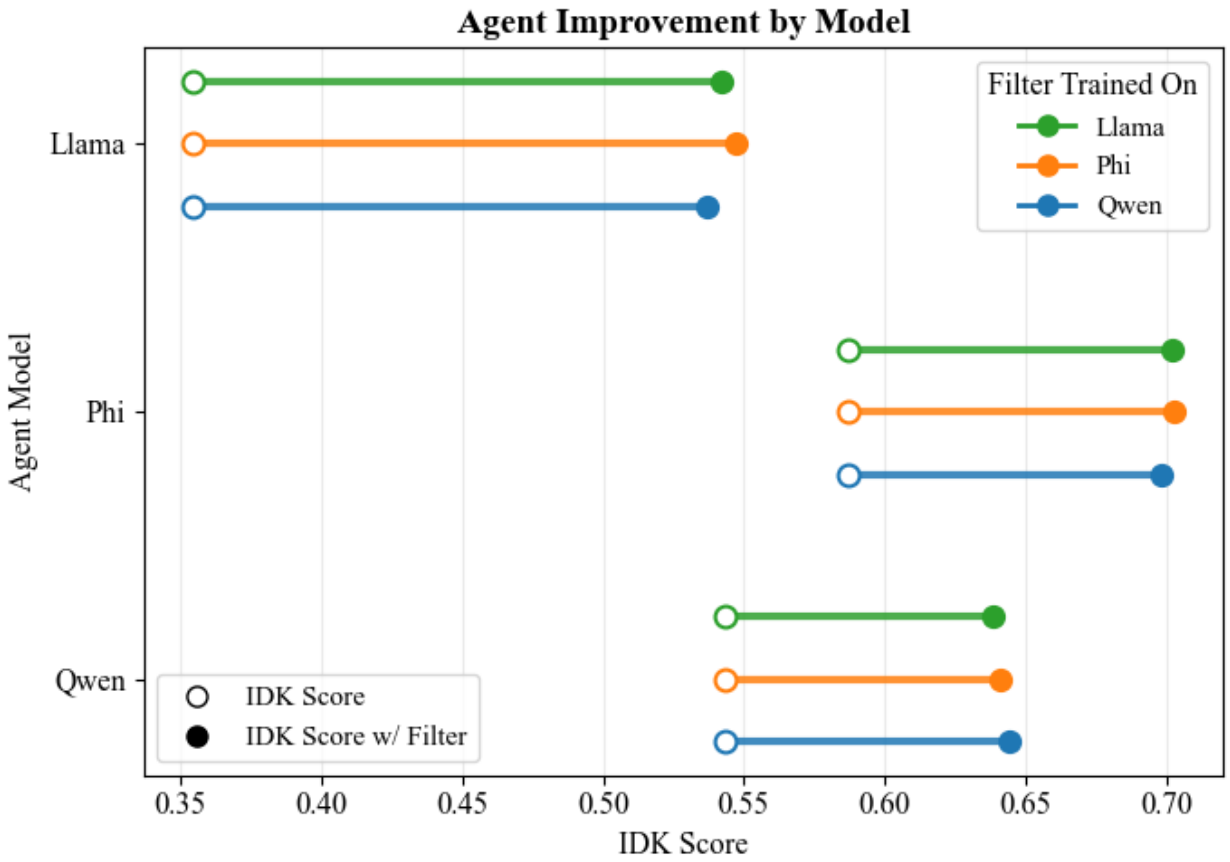


FIGURE 4.5. IDK Score for Llama, Phi, and Qwen agents averaged across all validation datasets, with and without a trained classifier acting as a filter. Results are grouped along the y-axis according to the agent model. Within a group, colors represent the model that was used to train the classifier filter. White dots represent the agent's performance without the filter and colored dots represent the agent's performance with a filter. Consistent performance gaps indicate that features are able to generalize across agent models.

and incorrectly treats the output as reliable. One observed example of this failure case is the agent being asked to find the average of a list of floats and producing a function call for a list of ints across all five repeated function calls. In such cases, the model's confidence is high but misplaced, and our filter has no signal with which to intervene. Addressing this failure mode likely requires complementary approaches, such as external type verification.

The second failure mode is superficial inconsistency due to argument formatting variation: when a model produces semantically identical function calls that differ only in surface-level formatting. For example, $3x\hat{}2$, $3 * x\hat{}2$, and $3 * x * *2$ may be treated as meaningful uncertainty by the

classifier and thus be incorrectly filtered out even if they all mean the same thing. One remedy would be to apply lightweight normalization to generated function calls before feature extraction (e.g., canonicalizing argument types), which we leave as a direction for future work.

4.6.3. Implications of Feature Invasiveness. The progressive improvement in IDKS as features move from blackbox to graybox to whitebox (Table 4.1) reflects a natural tradeoff between deployment practicality and classification performance. For practitioners working with API-gated models, our results show that blackbox features alone still yield meaningful gains over the baseline, and that logprobs, where available, provide a meaningful additional signal. Full whitebox access provides the strongest performance, but the marginal gains over blackbox and graybox features are modest on most benchmarks, suggesting that the core uncertainty signal is largely captured at the output and probability levels.

4.7. Conclusions

We have presented a framework with three main contributions: IDKS, a trained classifier filter, and validation across benchmarks and model families. These results matter because the deployment gap between benchmark performance and production reliability remains one of the primary barriers to adopting agents in high-stakes settings (Pan et al., 2026). By explicitly modeling and acting on uncertainty, encouraging agents to say "I don't know" rather than guess, our approach targets this gap directly. The framework is also intentionally lightweight, making it amenable to use in production settings. Notably, the classifier is a random forest trained on a modest feature set, and no modification to the underlying LM is required. Further, our use of a generalizable synthetic data generation pipeline demonstrates the feasibility of our approach even in settings where labeled data is sparse.

Several directions remain open for future work. First, principled threshold selection for the classifier is an important practical question that we have not extensively addressed here. The tradeoff between precision and recall in the IDK decision is one that will likely need to be calibrated per-domain. Second, the failure mode of confident incorrectness points to the limits of consistency as a primary proxy for uncertainty. Hybrid approaches that target other forms of agentic reliability (robustness, safety, and predictability) may help address this. Finally, while this work focuses on

SLMs with local access, extending the framework to closed-source LLMs is a natural next step as agents increasingly rely on frontier models in production.

Acknowledgement

Stefan Broecker and Thomas Strohmer would like to acknowledge support from the following grants: NSF DMS-2208356, NIH R01HL16351, P41EB032840, and DE-SC0023490.

The authors would like to acknowledge Zachary del Rosario for providing invaluable feedback on an early version of this work.

CHAPTER 5

Forgetting-MarI: LLM Unlearning via Marginal Information Regularization

This chapter is based on a paper that is currently under review and is a collaboration with Shizhou Xu from the UC Davis Department of Mathematics (at the time of the work), Yuan Ni from the SLAC National Accelerator Laboratory at Stanford University, and Thomas Strohmer from the UC Davis Department of Mathematics. It addresses a still-open question: how will LLM providers respond to an evolving regulatory and legal landscape governing what training data they can and cannot use? This chapter offers a mathematically rigorous, verifiable technique for removing information from a trained LLM. Unlearning is only one piece of responsible data management for LLM providers, but verifiable removal is a critical part of it. Together with the uncertainty quantification in Chapter 4, unlearning is essential to building and maintaining trust in LLMs as they become more embedded in everyday life.

5.1. Abstract

As Large Language Models (LLMs) face increasing regulatory scrutiny, the ability to surgically remove the influence of specific data without full retraining is critical, especially for systems that continuously accumulate unlearning requests. However, current LLM unlearning techniques are largely heuristic, lack formal guarantees, and often degrade model utility by suppressing information that is not unique to the unlearn set. We bridge the gap between rigorous unlearning theory and LLM practice by introducing *Forgetting-MarI*. This framework isolates and removes only the *marginal information*: the unique effect contributed by the unlearn set to the model's token-level predictive distribution, while preserving information supported by retained data. By penalizing marginal information, we derive a tractable upper bound on the residual influence of the unlearn set in the unlearned model, yielding a verifiable notion of behavioral undetectability against

confidence- and perplexity-based detection. Experiments on Llama and GPT models, including models up to 8B parameters, show that Forgetting-MarI is the only evaluated method that consistently reduces unlearn-set accuracy while preserving retain-set and validation performance, with particularly strong stability in continual unlearning. These results position marginal-information regularization as a principled and practical primitive for more controllable, auditable, and safe unlearning in real-world LLM deployments.

5.2. Introduction

As Large Language Models (LLMs) are increasingly deployed in high-stakes settings, selectively removing the influence of specific data has become critical. This need is driven by privacy regulations such as the GDPR "right to be forgotten," as well as practical concerns around copyrighted content, personally identifiable information, and harmful or biased data (Metz, 2024; Grynbaum et al., 2023; Freeman et al., 2024; Cyphert, 2023; Carlini et al., 2021). Machine unlearning seeks to remove a specified data subset's *influence* after training, avoiding the prohibitive cost of retraining from scratch.

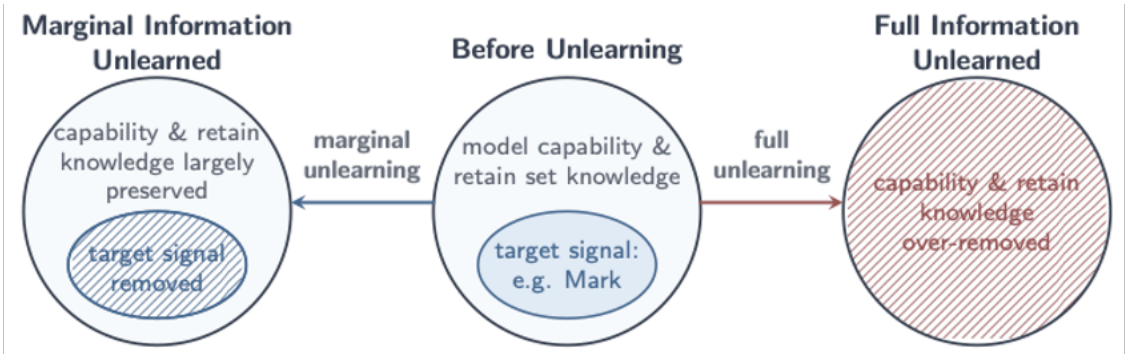
A central challenge is that existing LLM unlearning methods often *over-unlearn*: they suppress all information associated with the unlearn set, including information also supported by authorized retain data. Consider, for example, a copyright unlearning scenario in which an LLM was trained on two news articles about the same event, one from *The Washington Post* and one from *The New York Times*, but only the former is legally authorized for use. The two articles may share factual content about the event, while differing in phrasing, narrative style, ordering, and editorial perspective.

In this setting, there are two conceptually different unlearning objectives:

- **Marginal-information unlearning:** remove only the stylistic elements, phrasing, and content uniquely contributed by the unauthorized *Times* article, while retaining factual content independently supported by the authorized *Washington Post* article.
- **Full-information unlearning:** erase all content associated with the *Times* article, including factual information that is also independently supported by the authorized article.

We argue that marginal-information unlearning is the more natural objective for LLM unlearning. The goal is not to eradicate all knowledge appearing in the unlearn dataset, but to remove only its *marginal effect*: the component of the model's predictive behavior not already supported by authorized retain data. In the example above, marginal-information unlearning can satisfy the data-removal requirement without discarding facts that are lawfully present in the retain data. Full-information unlearning, in contrast, may unnecessarily remove shared knowledge and thereby degrade unrelated model capabilities.

Figure 5.1 provides a qualitative illustration of this distinction using Llama-3.2-1B completions.



Prompt	Seeing few other options, I casually switch out
Ground Truth to Forget	Mark's name card with that of a minor president on a better table.
Before Unlearning	Mark's name card with that of the president of Mexico
Marginal Information Unlearned	my Facebook profile picture and cover photo
Full Information Unlearned	`"types of@exmaple@example.mp accounts" for "types of.mp.mp.mp`

FIGURE 5.1. Comparison of sentence completions by Llama-3.2-1B. Before unlearning, the model memorizes the target text. *Full-information unlearning*, instantiated here by a gradient-ascent-based method, removes the target signal but also destroys linguistic utility. In contrast, *marginal-information unlearning* produces coherent completions that differ from the target, removing instance-specific information while preserving general language capability.

This distinction motivates our method, *Forgetting-MarI*, which directly targets the *marginal information*[1] of the unlearn set. Instead of erasing all information in $\mathcal{D}_u$, Forgetting-MarI suppresses

[1]Marginal information is formalized in Definition 5.2.1.

only the additional information contributed by $\mathcal{D}_u$ beyond what is already supported by the retain set $\mathcal{D}_r$. The utility term is therefore used to stabilize retained behavior, not to counteract an overly destructive unlearning objective.

In contrast, many existing LLM unlearning methods are *full-information* in principle: their ascent or preference losses target the entire signal of $\mathcal{D}_u$ (e.g., by maximizing CE on $\mathcal{D}_u$), while shared or legitimate knowledge is protected only indirectly through a competing retain loss, preference shaping, parameter subtraction, or orthogonality constraints. This creates an intrinsic tension between utility preservation and unlearning, often causing instability and requiring extensive tuning. Forgetting-MarI also differs from counterfactual and surrogate-target methods (Eldan et al., 2024; Ishibashi et al., 2024; S. Shi et al., 2024; Zheyuan Liu et al., 2024): rather than assuming manually designed alternatives or generic replacement responses, it iteratively samples from the retain set[2] to estimate and penalize marginal information.

5.2.1. Open Challenges in LLM Unlearning. Effective LLM unlearning must jointly address unlearning efficacy, utility preservation, and computational cost (S. Liu et al., 2025). Existing approaches remain limited along these axes. Full-parameter methods such as loss reversal, gradient difference, KL-ascent, and preference-based DPO/NPO often suppress $\mathcal{D}_u$ too broadly, degrading retain and held-out performance (B. Liu et al., 2022; Yuanshun Yao et al., 2024; Hong et al., 2024; Rafailov et al., 2023; R. Zhang et al., 2024a). Local editing methods such as ROME and MEMIT are efficient but struggle with distributed or stylistic information (Meng, Bau, et al., 2022; Meng, Sen Sharma, et al., 2023). Continual unlearning is also unresolved: exact retraining and shard-based retraining schemes provide strong removal but scale poorly with model size and request count (Cao et al., 2015; Ginart et al., 2019; Bourtoule et al., 2021), while inexpensive alternatives such as influence-function updates, repeated local edits, task-vector subtraction, model merging, or adapter-based patches can accumulate drift in retained behavior, interference between edits, or relearning across sequential requests (Guo et al., 2019; Meng, Bau, et al., 2022; Meng, Sen Sharma, et al., 2023; Ilharco et al., 2022; C. Gao et al., 2024; H. S. Kim et al., 2025; H. Xu et al., 2025). Finally, although certified unlearning is well established for simpler model classes, including linear

[2]We consider the standard supervised unlearning setting in which the unlearn set $\mathcal{D}_u$ is specified and the practitioner has access to an authorized retain data $\mathcal{D}_r$, or to a data stream serving the same role.

Method	Utility	Continual	Guarantees
Full Parameter	☆	✓	✗
Weight Editing	☆	✓	✗
Counterfactual	☆	✓	✗
Adaptation	☆	☆	✗
Ours	✓	✓	✓

TABLE 5.1. Summary of major LLM unlearning families along three open challenges: utility preservation, continual unlearning, and formal guarantees. Symbols: ✓=yes, ✗=no, ☆=partial.

models, kernel methods, and high-dimensional classifiers (Guo et al., 2019; Zou et al., 2025; S. Xu et al., 2025), extending comparable guarantees to billion-parameter autoregressive LLMs such as GPT and Llama remains a major open challenge. Table 5.1 summarizes these limitations across major method families.

5.2.2. Our Contributions. We introduce *Forgetting-MarI*, an information-theoretic framework for LLM unlearning based on the following conceptual notion (with a formal quantification in Definition 5.3.1).

DEFINITION 5.2.1 (Marginal Information (MarI)). *Marginal information is the additional effect on the model's token-level predictive distribution caused by adding the unlearn set to the retain set.*

Forgetting-MarI estimates and penalizes this marginal information, thereby removing the unique contribution of the unlearn set while preserving information supported by retained data. Our contributions are:

- **(A1) Conceptual Framework:** We formalize *Marginal Information Unlearning*, a paradigm shift that targets the surgical removal of unique data contributions (e.g., specific user traces), resolving the intrinsic conflict between forgetting and utility found in existing methods.
- **(A2) Theoretical Guarantees:** We derive an explicit upper bound on the residual mutual information between the model and the unlearn set. This provides the first provable unlearning guarantee (in terms of undetectability) for LLM unlearning. Such certification

Symbol	Meaning
V	vocabulary
T	sequence length
$\mathcal{D}_r, \mathcal{D}_u$	retain and unlearn datasets
$r \subseteq \mathcal{D}_r,\ u \subseteq \mathcal{D}_u$	retain/unlearn subsets, both contained in V^T
$d = r \cup u$	union subset used for MarI construction
μ_s	uniform law on subset $s \subseteq V^T$
$(p_\theta)_t^s(v)$	averaged next-token marginal on subset $s \in \{r, u, d\}$ at position t for token $v \in V$
p^r, p^u, p^d	collections of marginals $\{(p_\theta)_t^s\}_{t=1}^T$, $s \in \{r, u, d\}$
α	mixture weight $\frac{\lvert r\rvert}{\lvert r\rvert+\lvert u\rvert}$
(X_{MarI}, Z)	pair of random variables in the binary detection construction
$\ell_{\mathrm{MarI}}(\theta, r, u)$	marginal-information loss

TABLE 5.2. Core notation used in Section 5.3.1.

of unlearning efficacy is highly desirable to bridge the "mismatch between legal expectation and technical reality" (Hutson et al., 2026).

- **(A3) Scalability & Continual Learning:** We demonstrate that our additive mutual-information regularizer integrates seamlessly with gradient-based fine-tuning, enabling stable continual unlearning, a critical capability for lifelong models or agents that must process deletion requests without catastrophic forgetting of their core reasoning abilities.
- **(A4) Empirical Validation:** Extensive experiments on Llama and GPT models validate our theory, showing that Forgetting-MarI achieves superior trade-offs between forgetting efficacy and utility preservation compared to state-of-the-art baselines.

5.3. Unlearning via Marginal Information

Forgetting-MarI relies on a novel quantification of *marginal information* that (i) vanishes when the unlearn set $\mathcal{D}_u$ adds *no new* information beyond the retain set $\mathcal{D}_r$, and (ii) increases as $\mathcal{D}_u$ contributes information absent from $\mathcal{D}_r$. We instantiate this idea using a mutual-information (MI) objective. A summary of notation is provided in Table 5.2

5.3.1. Quantifying and Unlearning Marginal Effects. Fix a language model p_θ (with parameter θ) over a finite vocabulary V and a length $T \geq 1$. For $y \in V^T$, let $p_\theta(\cdot \mid y_{<t})$ be the next-token distribution. For a subset $s \subseteq V^T$, let μ_s be the uniform law on s and define

its averaged next-token marginals $(p_\theta)^s_t(v) := \mathbb{E}_{Y\sim\mu_s}\big[p_\theta(v \mid Y_{<t})\big]$ for $t \in [T]$, $v \in V$. Write $p^r := \{(p_\theta)^r_t\}_{t\in[T]}$, $p^u := \{(p_\theta)^u_t\}_{t\in[T]}$. For $d := r \cup u$, $p^d_t = \alpha\, p^r_t + (1-\alpha)\, p^u_t$, $\alpha := \frac{|r|}{|r|+|u|} \in (0,1)$. Let $T^* \sim \text{Uniform}([T])$ and $Z \sim \text{Bernoulli}(\frac{1}{2})$ be independent. Conditioned on $(T^* = t, Z)$, draw $X \sim p^d_t$ if $Z = 0$ and $X \sim p^r_t$ if $Z = 1$, and set $X_{\text{MarI}} := (T^*, X)$. Then the *mutual information* between X_{MarI} and Z is defined as

$$I(X_{\text{MarI}}; Z) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \text{JSD}\big(p^d_t, p^r_t\big). \tag{5.1}$$

Here, we denote the Jensen-Shannon divergence as $\text{JSD}(p,q) := \frac{1}{2}\,\text{D}_{\text{KL}}(p\|m) + \frac{1}{2}\,\text{D}_{\text{KL}}(q\|m)$ with $m := \frac{p+q}{2}$ and $\text{D}_{\text{KL}}(p\|q) := \sum_v p(v)\log\frac{p(v)}{q(v)}$. By construction, the information or distribution represented by d can be decomposed into the contribution of $r \cap d = r$ and the marginal contribution of $d \setminus r = u$. The distribution contributed by r through the model p_θ is p^r. The distributional contribution from the addition of u through p_θ is the distributional difference between p^r and p^d.

By construction, the quantification of the marginal effect is small if p^r is close to p^d, because such proximity suggests that the information content in u has already been largely represented by r. Conversely, the quantification will be large if p^r differs significantly from p^d, indicating that u contributes substantial new information w.r.t. p_θ and induces a model output distribution shift. Therefore, defining this marginal effect quantification boils down to differentiating p^r from p^d for any $r \subset \mathcal{D}_r$ and $d = r \cup u$ with arbitrary $u \subset \mathcal{D}_u$.

A natural way to *quantify* this difference is via a binary detection problem. Consider a binary detection problem using the construction above:

$$X_t := X\big|_{T^*=t} \sim \begin{cases} p^d_t, & Z = 0, \\ p^r_t, & Z = 1, \end{cases} \tag{5.2}$$

$\mathbb{P}[Z=0] = \mathbb{P}[Z=1] = \frac{1}{2}$. If $p^r = p^d$, even an optimal classifier does no better than a coin flip. If there is distributional shift, it can detect the difference. A sharp information–theoretic upper bound on the Bayes accuracy, denoted by P_{acc} and defined below in Proposition 5.3.1, is the following:

PROPOSITION 5.3.1 (Detection accuracy upper bounded by mutual information). *For* (X_{MarI}, Z) *with prior* $\pi = \mathbb{P}[Z = 1]$,

$$P_{\text{acc}} \;=\; \mathbb{E}\left[\max\{P(Z=0\,|\,X_{\text{MarI}}),\, P(Z=1\,|\,X_{\text{MarI}})\}\right] \;\le\; 1 - H_2^{-1}\big(H_2(\pi) - I(X_{\text{MarI}}; Z)\big),$$

where $H_2(\cdot)$ *is the binary entropy and* H_2^{-1} *denotes the inverse of* H_2 *restricted to* $[0, \frac{1}{2}]$.

Proof in Appendix C.1. Here $P(Z \,|\, X_{\text{MarI}})$ denotes the Bayes-optimal posterior between retain r and union d. Note that $I(X_{\text{MarI}}; Z) \in [0, H_2(\pi)]$ satisfies: (i) $I(X_{\text{MarI}}; Z) = 0$ when $p^d = p^r$; (ii) $I(X_{\text{MarI}}; Z)$ grows with their divergence, approaching $H_2(\pi)$ as $\mathcal{D}_r$ and $\mathcal{D}_u$ has mutually exclusive support. Since $p^d \neq p^r$ occurs precisely when u induces model confidence shifts not explained by r, Proposition 5.3.1 gives $I(X_{\text{MarI}}; Z)$ an intuitive meaning as the detectability of the marginal effect (Definition 5.2.1).

DEFINITION 5.3.1 (MI-based marginal information loss). *With* (X_{MarI}, Z) *as in* (5.2), *define*

$$\ell_{\text{MarI}}(\theta, r, u) \;:=\; I(X_{\text{MarI}}; Z).$$

Thus, Forgetting-MarI solves

$$\min_{\theta} \;\; \ell_{KL}(\theta, r) \;+\; \ell_{\text{MarI}}(\theta, r, u), \tag{5.3}$$

where $\ell_{KL}(\theta, r) := \mathrm{D}_{\mathrm{KL}}\big(p^r(\theta) \,\|\, p^r(\theta_0)\big)$ is the KL divergence between the updated model (parameter θ) and the frozen original model (parameter θ_0) on r, and ℓ_{MarI} is as above, motivated by Prop. 5.3.1. Algorithm 1 describes an efficient LLM implementation.

REMARK 5.3.1 (Alternative quantification). *Marginal information measures the shift from* $p_\theta(r)$ *to* $p_\theta(d)$. *Alternatively, one may use* $\ell'_{\text{MarI}}(\theta, r, u) := \mathrm{D}_{\mathrm{KL}}\big(p^d \,\|\, p^r\big)$ *or* $\mathrm{D}_{\mathrm{KL}}\big(p^d \,\|\, p^r(\theta_0)\big)$. *But mutual information has the advantage of (1) stability (boundedness), (2) interpretability (Proposition 5.3.1), and (3) continuous unlearning (evolving reference* $m = \frac{p+q}{2}$*).*

5.3.2. Marginal Information & Perplexity-Based Detectors. The current study of copyrighted text detectors can be roughly separated into two lines of work:

Algorithm 1 Forgetting-MarI

Require: Retain dataset $\mathcal{D}_r = \{x_i^r\}_{i=1}^{|\mathcal{D}_r|}$, unlearn corpus $\mathcal{D}_u = \{x_j^u\}_{j=1}^{|\mathcal{D}_u|}$; pretrained model f_{θ_0}; learning rate η; batch sizes B_r, B_u; epochs E; trade-off $\gamma \in (0,1)$; hetero $\in \{\texttt{False}, \texttt{True}\}$.
Ensure: Unlearned parameters θ_E.
Initialize $\theta \leftarrow \theta_0$; build dataloaders $\mathcal{L}_r, \mathcal{L}_u$.
for $e = 1$ **to** E **do**
Shuffle $\mathcal{L}_r, \mathcal{L}_u$; $S \leftarrow \min\{|\mathcal{L}_r|, |\mathcal{L}_u|\}$.
for $s = 1$ **to** S **do**
Minibatches: $r \leftarrow \mathcal{L}_r(s), \quad u \leftarrow \mathcal{L}_u(s)$.
Model logits & probabilities:
$L^r \leftarrow f_\theta(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p^r_{b,t} \leftarrow \text{softmax}(L^r[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
$L^r_0 \leftarrow f_{\theta_0}(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p^r_{b,t}(\theta_0) \leftarrow \text{softmax}(L^r_0[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
$L^u \leftarrow f_\theta(u) \in \mathbb{R}^{B_u \times T \times |V|}$, $p^u_{b,t} \leftarrow \text{softmax}(L^u[b,t,:]) \in [0,1]^{B_u \times T \times |V|}$;
$p^d \leftarrow \text{concat}_{\text{batch}}([p^r, p^u]) \in [0,1]^{(B_r+B_u) \times T \times |V|}$.
Utility Loss:
If hetero=`True`:
$\overline{p}^r \leftarrow \frac{1}{B_r * T} \sum_{b,t} p^r_{b,t}$, $\overline{p}^r_0 \leftarrow \frac{1}{B_r * T} \sum_{b,t} p^r_{b,t}(\theta_0)$, $\quad \ell_{\text{KL}} \leftarrow \sum_{v \in V} \overline{p}^r(v) \log(\frac{\overline{p}^r(v)}{\overline{p}^r_0(v)})$.
$\ell_{\text{KL}} \leftarrow \sum_{v \in V} \overline{p}^r(v) \log(\frac{\overline{p}^r(v)}{\overline{p}^r_0(v)})$
Else:
$p^r_t \leftarrow \frac{1}{B_r} \sum_{b=1}^{B_r} p^r_{b,t}$, $p^r_t(\theta_0) \leftarrow \frac{1}{N_r} \sum_{b=1}^{B_r} p^r_{b,t}(\theta_0)$,
$\ell_{\text{KL}} \leftarrow \frac{1}{T} \sum_t \big(\sum_{v \in V} p^r_t(v) \log(\frac{p^r_t(v)}{p^r_t(\theta_0)(v)}) \big)$
MarI Loss:
If hetero=`True`:
$P(v,0) \leftarrow \frac{1}{2(B_r+B_u)*T} \sum_{b,t} p^d_{b,t}$, $P(v,1) \leftarrow \frac{1}{2B_r*T} \sum_{b,t} p^r_{b,t}$, $P(v) = \sum_z P(v,z)$,
$\ell_{\text{MarI}} \leftarrow \sum_{z \in \{0,1\}} \sum_{v \in V} P(v,z) \log \frac{P(v,z)}{P(v)\,P(z)}$.
Else:
$P_t(v,0) \leftarrow \frac{1}{2(B_r+B_u)} \sum_b p^d_{b,t}$, $P_t(v,1) \leftarrow \frac{1}{2B_r} \sum_b p^r_{b,t}$, $P_t(v) = \sum_z P_t(v,z)$,
$\ell_{\text{MarI}} \leftarrow \frac{1}{T} \sum_t \big(\sum_{z \in \{0,1\}} \sum_{v \in V} P_t(v,z) \log \frac{P_t(v,z)}{P_t(v)\,P_t(z)} \big)$.
Total loss: $\mathcal{L} = (1-\gamma)\,\ell_{\text{KL}} + \gamma\,\ell^{\text{flat}}_{\text{MarI}}$.
Update: $\theta \leftarrow \theta - \eta\,\nabla_\theta \mathcal{L}$.
end for
end for
return $\theta_E \leftarrow \theta$

- **White-box methods:** These methods largely share the same idea of constructing a statistic (or a vector of statistics) that indicates the probability that a model has seen a given sentence or not. It bases the probability on how confidently the model predicts the true output. The idea is based on the intuition that a model that has seen the sentence during training will have high confidence when trying to complete it. Versions of this method include: minimum k-percent (W. Shi et al., 2024), perplexity outlier and reference

model perplexity outlier (Carlini et al., 2021), domain normalized minimum k-percentage (W. Zhang et al., 2024), and data-set level inference (Puerto et al., 2025).

- **Black-box methods:** These methods do not have access to the model parameters or the output logits or prediction distributions. They often use either edit distance (a.k.a. Levenshtein distance) or some token embedding model (e.g. a small transformer) to quantify the distance or similarity between a model's output and a reference string, then generate statistics of the similarity between the two. Examples of black-box detectors include direct regurgitation probes (Karamolegkou et al., 2023), Name-cloze membership inference (Fu et al., 2023), DE-COP: multi-choice preference (Duarte et al., 2024), and Output-consistency measures (Hu et al., 2025).

We provide theoretical guarantees for the unlearning performance of Forgetting-MarI against white-box copyright detectors that rely on model confidence (perplexity / cross-entropy). Let

$$S_\theta(x, y) \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(-\log p_\theta(x_t \mid y_{<t})\big)$$

be the standard cross-entropy (per-token negative log-likelihood). State-of-the-art detectors (Carlini et al., 2021; W. Zhang et al., 2024; Puerto et al., 2025) flag the membership of x in training by testing whether $S_\theta(x, x)$ is suspiciously low. We adopt the notation from Section 5.3.1: sequences $r, u \in V^T$, next-token marginals p_t^r, p_t^u, their mixture $p_t^d = \alpha\, p_t^r + (1-\alpha)\, p_t^u$ with $\alpha = \frac{|r|}{|r|+|u|}$, and the mutual information $I(X_{\text{MarI}}; Z)$.

The following theorem shows that small MarI makes the model's confidence on the unlearn sequence u explainable by retained/background information. Specifically, if $p^{\,d}$ is close to $p^{\,r}$ in the MarI sense, then the self-score $S_\theta(u, u)$ is close to the retain-context score $S_\theta(u, r)$.

THEOREM 5.3.2 (MarI controls the self-perplexity gap). *Fix* $u = (u_1, \ldots, u_T) \in V^T$ *and assume the* pathwise *non-vanishing condition* $\min\{p_t^u(u_t),\, p_t^r(u_t)\} \geq \gamma \in (0, 1]$ *for all* t. *Then*

$$\Big|\, S_\theta(u, u) - S_\theta(u, r) \,\Big| \;\leq\; \frac{2\sqrt{2}}{\gamma(1-\alpha)}\, \sqrt{I(X_{\text{MarI}}; Z)}\,.$$

See Appendix C.3 for the proof and Section 5.5.4 for numerical verification against state-of-the-art white-box detectors. Intuitively, as $I(X_{\text{MarI}}; Z) \to 0$, the self-perplexity gap vanishes. As a result,

confidence- or perplexity-based detectors lose discriminative power: high confidence on u can no longer be attributed uniquely to memorization of u.

We also note that the guarantee is distributional. The retain subset r need not be the original retain set itself, but only a sample from an authorized retained/background distribution against which the marginal contribution of u is measured. This distribution may be instantiated by the available retain set, an authorized proxy corpus, or a data stream serving the same role. Importantly, MarI is not merely sentence-level: $I(X_{\text{MarI}}; Z)$ averages position-wise divergences between p_t^d and p_t^r, so unlearn-set information expressed through elevated or shifted confidence at specific token positions also contributes to the loss. Hence, under the pathwise non-vanishing assumption, small MarI controls both sequence-level self-perplexity and token-position-specific confidence shifts.

REMARK 5.3.2 (Conditional information unlearning). *This work focuses on the unconditional MarI formulation, where marginal information is estimated over the sampled retain and unlearn sequences. The same principle can be extended to conditional unlearning by sampling prompts or contexts that encode the conditioning event. In that case, p_t^r and p_t^d become prompt-conditional next-token marginals, and small MarI removes the distinguishable contribution of the unlearn set under the sampled prompt distribution. A systematic treatment of conditional MarI is left for future work, as it requires specifying and covering the relevant prompt distributions.*

5.4. Algorithm Design

Sentence-wise MarI (Equation 5.1) compares p_t^d and p_t^r at each position t, preserving token-position structure. It therefore supports the stronger self-perplexity guarantee in Theorem 5.3.2, including control of position-specific confidence shifts. This estimator is most informative for homogeneous, context-aligned sentences in r and u, but can be noisy under heterogeneous batching. To improve stability, we also introduce a *pooled* or "flattened" estimator. It first averages over token positions and batch elements, $\overline{p}^{\,s} = \frac{1}{T}\sum_t p_t^s$, $s \in \{r, u, d\}$, and then computes $I(\overline{X}_{\text{MarI}}; Z) = \text{JSD}\left(\overline{p}^{\,d}, \overline{p}^{\,r}\right)$. Pooling filters position-heterogeneous noise and emphasizes dominant token-distribution shifts. Since it averages out position information, it yields a weaker but more stable word-level guarantee rather than the position-specific guarantee of sentence-wise MarI.

By the data-processing inequality, $I(\overline{X}_{\mathrm{MarI}}; Z)$ is a variational lower bound on sentence-wise MarI. The gap is controlled by the ℓ^2-deviation of token-sequence densities: see Theorem C.4.1 in Appendix C. Furthermore, pooled MarI also yields the following word-level guarantee:

THEOREM 5.4.1 (Word-level provable unlearning via pooled MarI). *Fix a token $w \in V$ and assume $\overline{p}^{\,u}(w) \wedge \overline{p}^{\,r}(w) =: \overline{\gamma}_w \in (0, 1]$. Then*

$$\left| \log \overline{p}^{\,u}(w) - \log \overline{p}^{\,r}(w) \right| \;\leq\; \frac{2\sqrt{2}}{\overline{\gamma}_w \,(1-\alpha)} \sqrt{I(\overline{X}_{\mathrm{MarI}}; Z)}\,.$$

This result is weaker than the sentence-wise guarantee in Theorem 5.3.2, because positional information has been averaged out. Nevertheless, it is sufficient for provable removal of word-level or token-frequency signals that do not depend on sentence structure.

The full algorithm is provided in Figure 1. In practice, we use the following protocol:

- **Estimator Selection:** Use *sentence-wise* MarI for homogeneous data (aligned contexts) to leverage precise signals. Use *pooled* MarI for large corpora with random batching to ensure stability.
- **Hyperparameter Tuning:** Fix a tolerable perplexity gap for the application, derive the required MarI threshold, and tune γ until the regularization magnitude falls below this threshold. This invokes the guarantees of Theorem 5.3.2 or 5.4.1.

Since both estimators achieve comparable empirical performance, Section 5.5 reports results under a unified label.

5.5. Experiments

Our experiments are designed to empirically validate the theoretical properties of Forgetting-MarI formalized in Section 5.3. Specifically, we address three key questions:

1 **Targeted Unlearning with Utility Preservation:** Does minimizing marginal information reduce unlearn-set accuracy toward the retrain-on-retain baseline while preserving retain-set, validation, and general benchmark performance?

2 **Continual Unlearning Stability:** Is the method robust to sequential unlearning requests (continual unlearning) without incurring the catastrophic drift or interference common in full-information methods?

3 **Provable Undetectability:** Consistent with Theorem 5.3.2, does the method defeat white-box perplexity detectors that rely on confidence scores to flag training data?

We compare against representative state-of-the-art full-parameter, optimization-based LLM unlearning baselines: **GA** (Yuanshun Yao et al., 2024), **GD** (B. Liu et al., 2022), **KL-GA** (Hong et al., 2024), and **DPO** (Rafailov et al., 2023). These methods cover the main full-information objective families in this setting: direct loss reversal, retain-counterbalanced optimization, KL-regularized retain preservation, and preference-based suppression. They give a controlled comparison with Forgetting-MarI, which is also implemented as a gradient-based update to the model parameters. The selected baselines cover the main objective families in this category: direct loss reversal on the unlearn set, retain-counterbalanced optimization, KL-regularized retain preservation, and preference-based suppression.

Although there are many variants of LLM unlearning, several recent methods remain within the same broader paradigm. NPO (R. Zhang et al., 2024b) uses a reference-model-based negative preference loss closely related to DPO-style unlearning. UNDIAL (Dong et al., 2025) distills the model toward manually adjusted logits, and PDU (Entesari et al., 2026) applies logit-flattening or entropy-increasing objectives on the unlearn set. These approaches differ in stabilization mechanisms, but they still suppress the full unlearn-set signal and rely on an auxiliary mechanism to preserve utility.

5.5.1. Experimental Setup.

5.5.1.1. *Protocol.* Following Maini et al. (2024) and Eldan et al. (2024), we adopt a rigorous three-step evaluation protocol to establish a theoretical ground truth: First, we fine-tune a base model on $\mathcal{D}_r \cup \mathcal{D}_u$, producing the Full Finetune model used as the starting point for all unlearning methods. Second, each method removes the influence of $\mathcal{D}_u$ from this model. Third, we train an Unlearn Baseline using only $\mathcal{D}_r$, never exposing it to $\mathcal{D}_u$. Successful unlearning should approach the Unlearn Baseline on $\mathcal{D}_u$ while preserving retain-set, validation, and general capability performance. We report next-token accuracy on $\mathcal{D}_r$, $\mathcal{D}_u$, and held-out validation sets, together with general capability metrics from the EleutherAI LM Evaluation Harness (L. Gao et al., 2024).

5.5.1.2. *Datasets & Models.* We evaluate Forgetting–MarI on two standard architectures: GPT-2 Large (774M) and Llama-3.2-1B. We select two text domains with distinct genres and pretraining

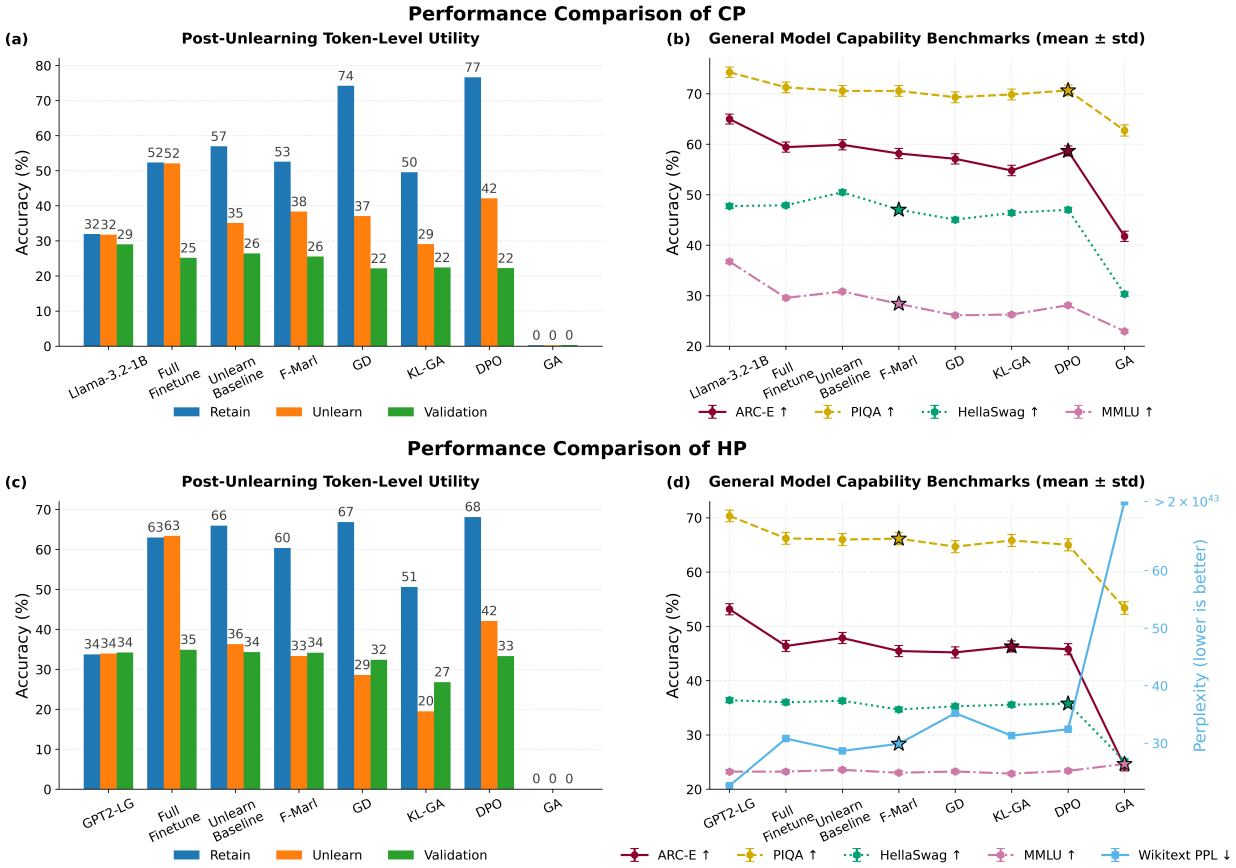


FIGURE 5.2. Targeted unlearning and utility preservation on *Careless People* (CP, top; Llama-3.2-1B) and *Harry Potter* (HP, bottom; GPT-2 Large). For each tuned method, the left panel reports token-level retain/unlearn/validation accuracy, while the corresponding right panel reports the same model's general capability benchmark performance. Each method is shown at its best hyperparameter and training epoch under the validation-preservation criterion. The goal is to reduce unlearn-set accuracy close to the Unlearn Baseline, i.e., the retrain-on-retain model, while largely preserving retain-set accuracy, validation accuracy, and benchmark performance. Forgetting-MarI is the only evaluated method that consistently achieves this behavior across both datasets. Stars indicate the best performer on each benchmark.

prevalence: (i) *Harry Potter and the Prisoner of Azkaban* (J.K. Rowling, 2013), a popular choice of unlearn set but likely to present in models' pretraining; and (ii) *Careless People: A Cautionary Tale of Power, Greed, and Lost Idealism* (Sarah, 2025), published after Llama's release and thus unlikely to appear in its pretraining.

5.5.1.3. *Splits.* For *Harry Potter* (HP), we designate 10% of sentences as $\mathcal{D}_u$ and the remaining 90% as $\mathcal{D}_r$ (cf. Eldan et al., 2024); validation uses excerpts from another book in the series (*Harry*

Potter and the Sorcerer's Stone (Rowling, 2023)). For *Careless People* (CP), we use a split with contiguous 50/50 spans for $\mathcal{D}_u/\mathcal{D}_r$ and use stories published on Reddit's WritingPrompts subreddit in 2025 (post-release) as validation.

5.5.1.4. *Parameter Tuning & Runtime Information.* For each method, we stop training once validation accuracy drops by more than 3% from its initial value (to prevent general-utility degradation).

5.5.2. Targeted Unlearning with Utility Preservation. The Unlearn Baseline denotes the retrain-on-retain model, trained only on $\mathcal{D}_r$ and never exposed to $\mathcal{D}_u$. It serves as the gold-standard target: a successful method should reduce unlearn-set accuracy toward this baseline while preserving retain-set, validation, and general benchmark performance.

5.5.2.1. *Token-Level Utility (Left Panels, Fig. 5.2).* Across both HP and CP, Forgetting-MarI is the only evaluated method that consistently moves unlearn-set accuracy toward the Unlearn Baseline while preserving retain-set and validation accuracy.

- *Careless People (Llama-3.2-1B):* Full-information baselines exhibit unstable behavior. GD and DPO substantially increase retain accuracy, indicating overfitting to $\mathcal{D}_r$ while failing to match the validation behavior of the Unlearn Baseline. KL-GA preserves less of the desired forgetting behavior and also degrades validation accuracy. Forgetting-MarI, by contrast, reduces unlearn-set accuracy while maintaining retain and validation performance close to the Unlearn Baseline.
- *Harry Potter (GPT-2 Large):* GD and DPO again fail to simultaneously lower unlearn-set accuracy and preserve validation accuracy, while KL-GA either under-unlearns or degrades validation performance. Forgetting-MarI is the only method that lowers unlearn-set accuracy close to the Unlearn Baseline without dropping validation accuracy.

5.5.2.2. *General Capability (Right Panels, Fig. 5.2).* We evaluate standard benchmarks using the EleutherAI LM Evaluation Harness: ARC-E, PIQA, HellaSwag, WikiText, and MMLU. The right panels show that Forgetting-MarI preserves general capability while achieving targeted unlearning on the left panels.

- *Llama-3.2-1B on CP:* Forgetting-MarI achieves the best performance on HellaSwag and MMLU and ranks second on PIQA and ARC-E. Although DPO obtains slightly higher scores on PIQA

and ARC-E, it does not reduce unlearn-set accuracy to the desired level. Thus, among methods that successfully unlearn, Forgetting-MarI best preserves general capability.

- *GPT-2 Large on HP:* Forgetting-MarI ranks first on PIQA and WikiText, second on ARC-E, and remains competitive on HellaSwag. This setting illustrates why benchmark accuracy alone is insufficient for evaluating unlearning: GA attains the highest MMLU score, but its WikiText perplexity deteriorates severely, showing that isolated benchmark gains can coexist with damage to the language-modeling distribution. Accordingly, we base our main conclusions on token-level and distribution-level behavior, including retain/unlearn/validation accuracy and perplexity-based measures, and use benchmark scores as supplementary evidence of preserved model capability.

Overall, Forgetting-MarI is the only evaluated method that lowers unlearn-set accuracy close to the Unlearn Baseline while largely preserving retain-set accuracy, validation accuracy, and general benchmark performance.

5.5.3. Continual Unlearning with Sequential Requests.

5.5.3.1. *Setup.* To simulate repeated deletion requests, we partition the total unlearn set into three disjoint subsets, $\mathcal{D}_u = \mathcal{D}_{u,1} \cup \mathcal{D}_{u,2} \cup \mathcal{D}_{u,3}$. Starting from the full model trained on $\mathcal{D}_r \cup \mathcal{D}_u$, we apply each method sequentially to remove one subset at a time.

- **Careless People (random split):** We randomly partition the unlearn set into three equal, disjoint subsets and remove them sequentially.
- **Harry Potter (character-based):** To emulate "Right to be Forgotten" user requests, we construct a "forget-characters" benchmark by assigning sentences to characters using alias matching (e.g., "Granger" $\rightarrow$ Hermione). We sequentially unlearn Hermione, then Snape, then Ron. The retain set contains all "other" sentences, all retain-character sentences (e.g., Harry and Dumbledore), and sentences associated with characters not yet targeted for unlearning.

5.5.3.2. *Results (Table 5.3).* Table 5.3 evaluates continual unlearning along three axes: stability of retain and validation accuracy, successful removal of the current target and persistence on previously targeted unlearn sets, and preservation of ARC performance as a common general-capability metric.

Low → High utility / ARC performance Over-unlearn → Under-unlearn

Setting			Utility		Unlearn Sets			General	
Data	Metho	Step	Ret.↑	Val.↑	F1/He	F2/Sn	F3/Ro	ARC↑	Status
CP: Llama	KL-GA	First 15%	43.1	16.3	7.5	32.2	34.0	53.7	Unstable
		Second 15%	31.1	5.3	3.0	1.8	12.0	53.2	Collapse
		Third 15%	21.9	2.4	1.5	1.1	0.9	52.1	Collapse
	DPO	First 15%	75.0	22.9	42.5	46.1	46.5	57.2	Under-unlearn
		Second 15%	82.6	20.8	40.4	38.3	42.8	54.9	Under-unlearn
		Third 15%	85.9	20.1	39.1	37.4	36.2	52.0	Under-unlearn
	GD	First 15%	76.5	19.3	21.3	41.4	41.4	53.8	Unstable
		Second 15%	74.7	12.7	15.4	7.0	30.2	48.9	Unstable
		Third 15%	76.5	8.6	11.5	5.8	3.8	41.5	Collapse
	F-MarI	First 15%	**51.9**	**25.3**	**35.0**	**49.7**	**50.0**	**58.6**	**Stable**
		Second 15%	**51.6**	**25.1**	**34.7**	**30.4**	**47.7**	**57.9**	**Stable**
		Third 15%	**51.2**	**25.0**	**34.3**	**30.2**	**27.8**	**55.9**	**Stable**
HP: GPT2-LG	KL-GA	Hermione	51.0	33.5	4.7	58.0	59.6	46.1	Unstable
		Snape	52.2	34.2	9.4	9.2	61.8	46.2	Relearn
		Ron	49.3	28.9	9.6	14.7	7.1	44.6	Relearn
	DPO	Hermione	67.3	33.3	44.6	69.0	69.5	46.4	Under-unlearn
		Snape	69.9	33.5	45.0	46.4	71.9	45.5	Under-unlearn
		Ron	79.9	33.4	42.0	46.0	50.4	46.2	Under-unlearn
	GD	Hermione	66.3	33.3	11.7	67.2	67.7	44.1	Unstable
		Snape	68.8	33.2	12.0	15.7	70.1	43.1	Unstable
		Ron	78.2	31.7	10.7	15.8	16.0	42.8	Collapse
	F-MarI	Hermione	**60.0**	**36.3**	**16.2**	60.9	62.9	**46.1**	**Stable**
		Snape	**60.3**	**36.2**	**15.5**	**14.0**	63.1	**45.2**	**Stable**
		Ron	**60.4**	**34.9**	**13.4**	**15.0**	**15.2**	**44.0**	**Stable**

TABLE 5.3. **Continual unlearning under sequential requests.** For each method, the three rows are the three successive unlearning steps. Retain, Validation, and ARC should remain high and stable across steps. For Unlearn Sets, lower values indicate stronger removal: diagonal entries correspond to the currently targeted set, and entries below the diagonal test whether previously targeted sets remain low after later requests. Relearning of previously targeted sets, unnecessary over-unlearning, and deterioration of retain/validation/general performance all indicate continual-unlearning instability. Forgetting-MarI is the only evaluated method that consistently keeps current and previous unlearn-set accuracies low while preserving retain, validation, and ARC performance.

- *Careless People with Llama-3.2-1B.* Forgetting-MarI is the only method that keeps retain and validation accuracy stable while maintaining low accuracy on current and previously targeted unlearn sets, without relearning or over-unlearning. It also preserves ARC performance across all three requests. By contrast, KL-GA over-unlearns and eventually collapses in validation accuracy, DPO leaves high unlearn-set accuracy, and GD overshoots on $\mathcal{D}_r$ while degrading validation and ARC performance.
- *Harry Potter with GPT-2 Large.* Forgetting-MarI again shows the most stable behavior: later requests do not relearn previously unlearned character sets, and retain, validation,

and ARC performance remain stable. KL-GA lowers accuracy on the current unlearn set but later relearns previous unlearn sets while degrading validation accuracy. GD lowers some unlearn-set accuracies but causes substantial validation or general-capability degradation, whereas DPO retains high unlearn-set accuracy, indicating under-unlearning.

5.5.3.3. *Conclusion.* Forgetting-MarI delivers robust sequential unlearning: it keeps current and previously targeted unlearn-set accuracies low, avoids relearning under later requests, and preserves retain, validation, and general benchmark performance more consistently than full-information baselines.

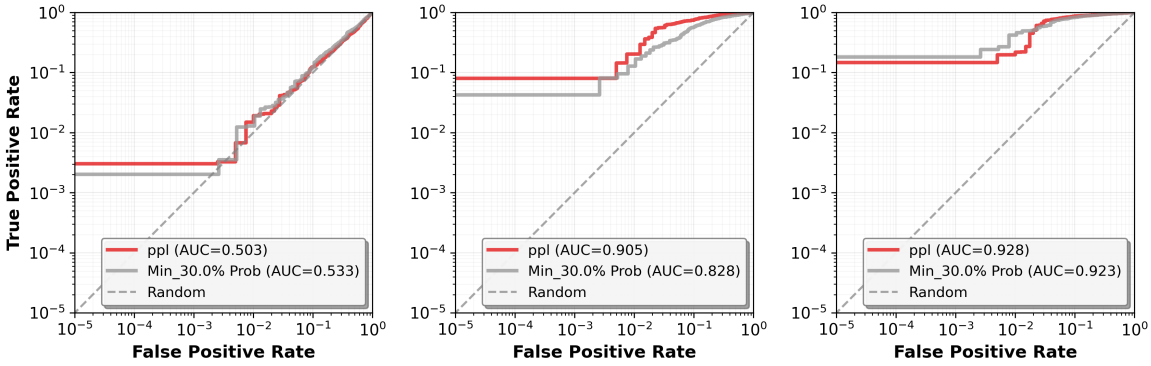


FIGURE 5.3. ROC curves for a state-of-the-art white-box membership inference detector. Left: Full Finetune model, where the unlearn set remains detectable. Middle: Forgetting-MarI. Right: Unlearn Baseline, trained only on $\mathcal{D}_r$. The similarity between Forgetting-MarI and the Unlearn Baseline empirically supports Theorem 5.3.2: minimizing marginal information reduces detectability of the unlearn data under this confidence-based white-box detector.

5.5.4. Detector Evaluation: Empirical Verification of Guarantees. Theorem 5.3.2 shows that minimizing $I(X_{\text{MarI}}; Z)$ controls the gap between the self-perplexity of the unlearn data and the perplexity induced by retained/background information. Therefore, confidence-based detectors that rely on perplexity, cross-entropy, or likelihood-ratio gaps should lose discriminative power when this marginal-information gap is small.

We evaluate this prediction using white-box detectors that access model logits (Carlini et al., 2021; W. Shi et al., 2024). In particular, we use the Min-K% detector (W. Shi et al., 2024), together with perplexity-based detection, and compare three models: (i) the Full Finetune model before unlearning, (ii) Forgetting-MarI, and (iii) the Unlearn Baseline trained only on $\mathcal{D}_r$.

We report ROC curves with the score oriented so that higher AUC corresponds to behavior closer to non-membership of $\mathcal{D}_u$. Under this convention, low AUC indicates that the detector still identifies the model as having seen $\mathcal{D}_u$, while high AUC indicates behavior closer to a model trained only on $\mathcal{D}_r$.

As shown in Figure 5.3, the Full Finetune model remains distinguishable from the non-member behavior, as expected before unlearning. In contrast, Forgetting-MarI achieves ROC curves close to the Unlearn Baseline for both perplexity and Min-K% detection. This supports the prediction that marginal-information minimization reduces confidence-based detectability of the unlearn data in this white-box detector setting.

5.6. Conclusion

This work introduces *Forgetting-MarI*, a paradigm shift in LLM unlearning that moves beyond the heuristic "full-information" removal with utility regularization common in current approaches. By formalizing the concept of Marginal Information, we resolve the intrinsic conflict between forgetting a target set $\mathcal{D}_u$ and preserving the utility of a retain set $\mathcal{D}_r$. Our framework provides the first rigorous theoretical guarantees for LLMs unlearning, proving that minimizing marginal information suffices the undetectability of the unlearned data.

Our contributions pave the road for a new generation of theoretically grounded unlearning methods. While empirical results on Llama and GPT models support that our approach matches the retrain-on-retain "Gold Standard", several critical avenues for research remain. Currently, our definition of marginal information relies on access to $\mathcal{D}_r$. Future work could substitute $\mathcal{D}_r$ with synthetic data or new, independent datasets. This would not only enable "zero-shot" unlearning in privacy-sensitive deployments where $\mathcal{D}_r$ is inaccessible but also improve fine-tuning robustness by preventing overfitting to a static retain set. Our theory relates λ to the classic rate-distortion in information theory. Future research could derive closed-form or learning-based solutions for optimal parameter selection, removing the need for validation loops.

5.7. Broader Impacts

This work studies machine unlearning for LLMs, with potential benefits for data-removal workflows, privacy- and copyright-sensitive model management, and more auditable model deployment.

By targeting marginal information rather than all information associated with the unlearn set, Forgetting-MarI is designed to reduce data-specific influence while preserving behavior supported by retained data.

The scope of these impacts is tied to the scope of the method and its assumptions. Our theoretical results are stated relative to the retained/background distribution and the token-level predictive behavior used to estimate marginal information. They should therefore be interpreted as guarantees within the specified distributional and detector setting, not as a universal certificate of removal for all possible adversarial prompts or attacks, especially directly on the model parameter space level.

CHAPTER 6

Estimating Temperature in LLM-Generated Text: A Maximum Likelihood Approach and Its Limits

This chapter is based on (as yet) unpublished work in collaboration with Thomas Strohmer from the UC Davis Department of Mathematics. It presents a technique for estimating the temperature used to generate text given just the piece of text. The work grew out of 1) a desire to understand language models more deeply while beginning the work that became Chapter 5, as well as 2) a yet-unanswered question about whether human-written text has any analog to LLM-generated text's temperature. By taking a more mathematically rigorous approach to model characterization than much of the surrounding literature, this work gives its conclusions firmer theoretical footing. Moreover, in exploring how different language models may encode language and meaning in subtly incompatible ways, this chapter closes the dissertation's arc by deepening our understanding of what these models are actually doing, and that deeper understanding informs the decisions we make about how much we should trust them.

6.1. Abstract

The randomness of token sampling in large language model (LLM) decoding is controlled by a hyperparameter called temperature and plays a central role in shaping the character of generated text. We introduce a method for estimating the temperature used to generate a given piece of text without requiring white-box access to the generating model, external classifier training, or large sample sizes. Our approach feeds observed text into a surrogate language model to obtain estimated next-token logit distributions, then computes the maximum likelihood estimate of temperature under those distributions. Because the method relies on standard maximum likelihood theory, we derive theoretical accuracy guarantees via the Cramér–Rao bound and Fisher information, and validate these bounds through both simulation and experiments with real language models.

When the surrogate model matches or shares a model family with the generating model, our estimator recovers temperature accurately and with error closely matching theoretical predictions. When the surrogate model comes from a different model family, however, estimation accuracy degrades in a consistent, S-shaped pattern: low temperatures are overestimated, high temperatures are underestimated, and a middle range is estimated reliably. We discuss possible explanations for this cross-family degradation, including logit distribution mismatch from divergent training processes and the effects of maximum likelihood estimation under model misspecification, and outline directions for future work to disentangle these causes. Our results offer a lightweight, theoretically grounded tool for characterizing LLM decoding behavior and surface open questions about how differently trained language models represent language.

6.2. Introduction

With the introduction of GPT-5, OpenAI has stopped supporting the use of *temperature* in its language model APIs when using newer models (Jim, 2025). While many other providers of large language model (LLM) still allow for users to customize temperature in their APIs, OpenAI's restriction is an acknowledgment of the importance of the parameter, which is often described as controlling the "creativity" of the model. Indeed, tuning a decoding algorithm to human preferences takes significant time, manual effort, and computation, so it's no surprise that model providers would want to control it, and thus control the user's experience with their model (Naseh et al., 2023). The resources required to tune a decoding algorithm also makes knowledge of the hyperparameters extremely valuable, and makes the inference of those decoding parameters an active area of study.

Beyond the potential financial value of the hyperparameters of a decoding algorithm, the growing ubiquity of LLM-generated text that gets encountered everyday makes characterization of decoding parameters an important question for researchers. Much of the research in this area focuses on distinguishing LLM-generated text from human-generated text, with numerous commercial (LangChain, 2026) and academic (Crothers et al., 2023) options available for use. One of the first successful demonstrations of detecting LLM-generated text was DetectGPT (Mitchell et al., 2023). Recent work has drawn attention to the effect that temperature can have on the accuracy of these LLM detectors. Schneider et al. (2023) trained classifiers for LLM detection and showed that

temperature can affect the accuracy of their method. Meanwhile, Kempton et al. (2025) developed a method that they claim to be robust to temperature changes during text generation.

For work that goes beyond LLM detection, focus has largely been on identifying the decoding algorithm itself and not the hyperparameters of that algorithm. Ippolito et al. (2023), for example, use repeated queries with differing prompts to distinguish between top-p and top-k sampling and to estimate p and k. Similarly, Tay et al. (2020) trained classifiers to predict the decoding strategy used to generate text.

This work focuses on the hyperparameters themselves, and specifically temperature. It introduces an algorithm for estimating the temperature that was used to generate a piece of text, given only the text. Of note for this work, Naseh et al. (2023) developed a method to estimate the sampling method used and various sampling parameters, including temperature. However, this work distinguishes itself from Naseh et al. (2023)'s work and others in a few key ways. Namely:

- No external model is trained.
- White-box access of the generating model is not required.
- Only text access is required.
- Temperature estimation can be performed with a single piece of text, as opposed to hundreds or thousands.
- A theoretical accuracy bound can be calculated.

In addition to describing the inference algorithm and providing experimental validation of its effectiveness, this work provides an exploration of the algorithm's failure modes and a discussion of possible explanations and implications for those failure modes.

6.3. Background and Approach

Very generally, LLMs answer the question: given a sequence of tokens, what token comes next? More precisely, they calculate the conditional probability of the next token given all preceding tokens:

$$P(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} P(w_t \mid w_1, w_2, \ldots, w_{t-1}), \tag{6.1}$$

where $w_1, w_2, \ldots, w_t$ is a sequence of T tokens. A key thing to note, however, is that language models themselves don't generate token probabilities, they generate logits. The process of going from logits to token generation is performed by a decoding algorithm. Techniques abound for decoding logits into token selections, but the most common are based on sampling, which will be the focus of this work. In sampling, the logits at each generation step are converted into a probability distribution that is then sampled from to select the next token. In it's "pure" form, logits are converted into probabilities using the softmax function conditioned on some *temperature*, τ. At each step $s \in [1, T]$, one token is chosen from a vocabulary of size K. The probability of selecting token i at step s with temperature τ is given by:

$$p_{is}(\tau) = \frac{e^{l_{is}/\tau}}{\sum_{j=1}^{K} e^{l_{js}/\tau}}, \tag{6.2}$$

where l_{is} represents the logit for token i at step s. This equation gives insight into why temperature is often called the "creativity" parameter. At high values of τ the probability distribution is more evenly spread among all tokens, increasingly the likelihood that a less common token is selected, and that a more "creative" sequence gets generated.

Variations on this pure method apply filters to the set of likely next tokens. Notable examples include top-k sampling, which only samples from the distribution that includes the k most likely tokens (Holtzman, Buys, Forbes, et al., 2018), and top-p sampling, which samples from the minimum tokens needed to cumulatively make up p% of the token probability weight (Holtzman, Buys, Du, et al., 2020).

Given the probability of selecting a token at each step from Equation 6.2, the likelihood of the complete sequence of observations X is:

$$P(X|\tau) = \prod_{s} p_{x_s,s}(\tau), \tag{6.3}$$

where x_s denotes the index of the chosen token at step s. An estimate of the temperature, $\hat{\tau}$, can then be found by maximizing the likelihood over the whole sequence:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\, P(X|\tau) \tag{6.4}$$

The key, then, is to get an accurate estimate of the logit distribution at each step of the generation process without access to the model that originally generated the text. Inspired by works that argue that representations across language models are converging as their performance improves (Huh et al., 2024), we use the logits of a language model that we do have access to to estimate the logits of the language model we do not have access to.

Given a piece of text, our approach to temperature estimation is therefore:

(1) Feed the text sequence into a surrogate, known language model, token by token

(2) For each token, record the logit distribution for the next predicted token.

(3) Calculate the value of τ that maximizes the likelihood of the known sequence being generated, given the estimated logit distribution at each step. Note that l_{is} in Equation 6.2 above is now estimated by the surrogate model rather than observed directly

6.4. Theoretical Analysis

Since our method is based on maximum likelihood estimation (MLE), we can make theoretical guarantees about the accuracy and error bounds of our temperature estimate. We then show that these guarantees hold empirically.

First, it has been shown that, under regularity conditions, maximum likelihood estimators are asymptotically efficient, meaning that they achieve the Cramér–Rao lower bound as $n \to \infty$, where n is the length of the token sequence (Gundersen, 2019). Our method uses the BFGS algorithm to approximate the estimator (Fletcher, 2000), and meets the following regularity conditions (Lehmann et al., 2006):

(1) *Parameter identifiability*: Different parameter values must lead to different probability distributions.

(2) *Support independence*: The set of possible observations shouldn't depend on the parameter value.

(3) *Differentiability*: The log-likelihood function needs to be twice differentiable with respect to the parameter.

(4) *Bounded information*: The Fisher information must be finite and positive.

(5) *Exchangeability of integration and differentiation*: Allows for switching the order of differentiation and expectation.

(1) holds because different temperature values produce distinct token distributions. Any token can be selected regardless of the temperature value (though with different probabilities) so (2) holds. We calculate the second derivative derivative below, so (3) holds. (4) holds as long as there's some variation in the logits and $\tau > 0$. (5) holds, and can be shown by simply performing the calculation, as in (3).

Since all of the regularity conditions hold, under correct specification, pure softmax sampling, and sufficient variation in the logits, the maximum likelihood estimator is consistent and asymptotically efficient. Further, because an efficient estimator achieves the Cramér-Rao lower bound, we can derive bounds on the variance of our estimator. Namely, the Cramér-Rao lower bound says that the variance of any unbiased estimator $\hat{\theta}$ of θ is bounded by the reciprocal of the Fisher information $I(\theta)$, or:

$$var(\hat{\theta}) \geq \frac{1}{I(\theta)}. \tag{6.5}$$

Fisher Information is the expected value of the square of the derivative of the log likelihood:

$$I(\tau) = \mathbb{E}\left[\left(\frac{\partial}{\partial \tau} \log P(X|\tau)\right)^2\right]. \tag{6.6}$$

For our method, this gives:

$$I(\tau) = \mathbb{E}\left[\left(\sum_s \frac{1}{\tau^2}(-l_{x_s,s} + W_s(\tau))\right)^2\right] = \frac{1}{\tau^4}\mathbb{E}\left[\left(\sum_s (-l_{x_s,s} + W_s(\tau))\right)^2\right], \tag{6.7}$$

where $W(\tau)$ at each step s is defined as:

$$W(\tau) = \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}. \tag{6.8}$$

A full derivation of the Fisher Information can be found in Appendix D.1.

6.5. Simulation-Based Validation

We can verify the analysis above using simulations. Specifically, we can simulate the generation of text with a toy model and empirically validate the accuracy and variance bounds of our estimator. The simulation looks roughly like this:

(1) Randomly generate a sequence of "logits" by sampling from a distribution, such as a normal distribution. Sample V logits for each of s steps, where V is the size of your simulated vocabulary and s is the length of your simulated token sequence.
(2) For each step, transform the logits into a probability distribution using a known temperature, τ.
(3) Sample from the distribution at each step to "generate" a token.

Figure 6.1 shows the average performance of our estimator when logits are sampled from a normal distribution. It shows that, as the sequence length increases, $\hat{\tau}$ approaches τ and the observed standard deviation of our estimator approaches the standard deviation we would expect based on Fisher Information.

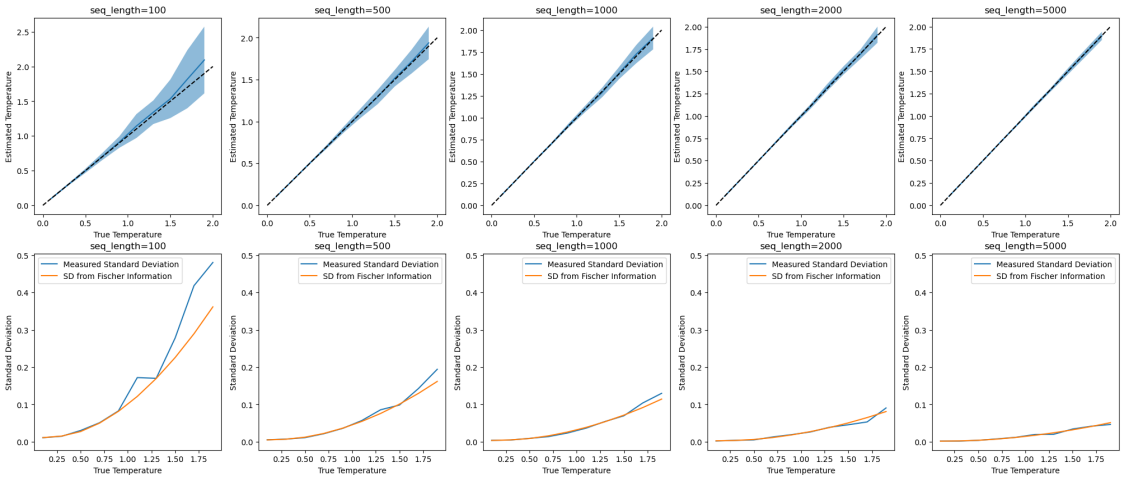
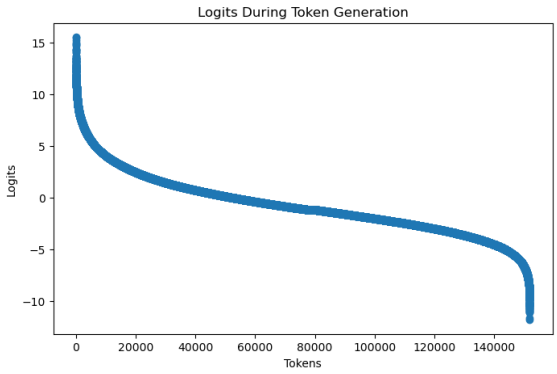


FIGURE 6.1. **Top row**: Plots showing the agreement between the true temperature and the estimated temperature in simulations with varying sequence lengths. Dotted lines show a perfect estimator, while blue lines show the true performance of the estimator. Blue shadows indicate standard error. **Bottom row**: Plots showing the agreement between theoretical standard deviation based on Fisher Information and observed standard deviation during simulations at different temperatures. Results at each temperature are averages based on 50 randomly generated token sequences.

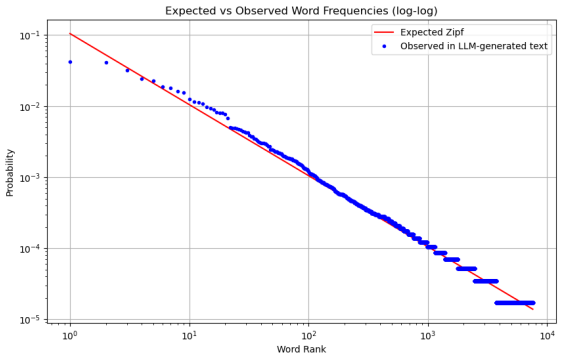
A basic assumption in this simulation is that logits are normally distributed, which is not the case in practice. Figure 6.2a shows a representative example of what logits actually look like during generation in a language model. Moreover, the simulation assumes that tokens are generated independently, which is also not the case in practice. The likelihood of a given token is dependent upon the tokens preceding it, and some tokens are generally much more likely than others. Indeed, language is often observed to follow a power law, where the frequency of a word in a vocabulary is inversely proportional to the rank of that word in the language (Zipf, 1935). This relationship can be seen in LLM-generated text as well, as shown in Figure 6.2b, where observed token frequencies closely match the expected frequencies predicted by a power law relationship.

6.6. Experiments with Language Models

6.6.1. Validating Simulations. To validate the results of the simulations in the previous section given the noted limitations of our approach, we performed the same experiment with real language models. Though, it should be noted, due to computational limitations, experiments with real language models are limited to relatively small versions, ranging from one to seven billion



(A) Raw logit values for all tokens in the Llama tokenizer's vocabulary during one step of token generation.

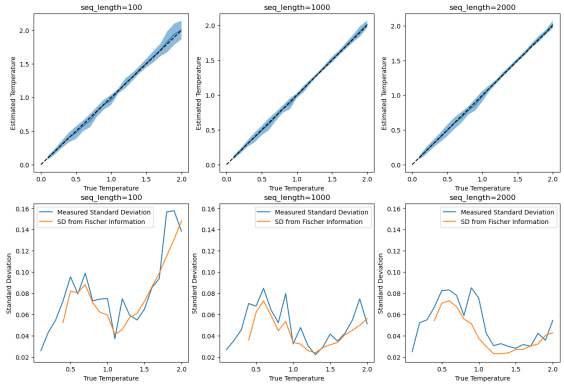


(B) Token frequency, sorted by frequency rank, of tokens generated by Llama-3.2-1B.

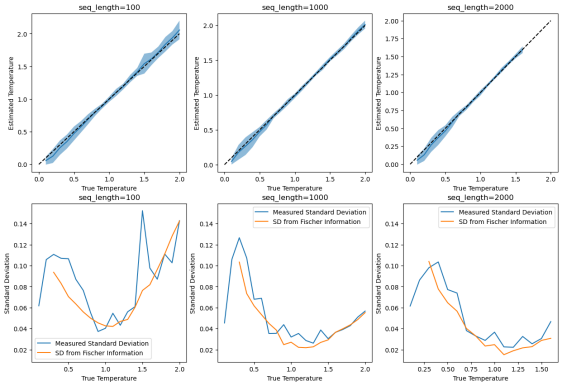
FIGURE 6.2. **Left**: Logit values observed during token generation in a language model. **Right**: The relationship between token rank and probability during language model generation.

parameters. In the experiments, language models are minimally prompted – given just a start token – and sequences of text are generated using basic sampling, where logits are transformed into a probability distribution using a softmax and no extra sampling techniques are used, such as top-p or top-k sampling. Logits are then generated using the same model as the surrogate language model, and those logits are used to estimate the temperature. The models used are Meta's Llama-3.2-1B and Alibaba Cloud's Qwen2.5-1.5B (A. Yang et al., 2024; Grattafiori et al., 2024). For each experiment, sequences of length 100, 1,000, and 2,000 tokens are generated at temperatures between 0.1 and 2.0 at increments of 0.1, with 10 repeats across each temperature and sequence length combination. Figure 6.3 shows the results of these experiments. Just like with simulated language models, our estimator is able to accurately estimate the temperature of generated text, and the observed error closely matches the theoretical error based on Fisher Information.

6.6.2. Measuring Temperature Across Models. Moving beyond a model being used to estimate its own temperature, we ran experiments where the generating model and the surrogate model differ. In the first set of experiments, we explore how our estimator works among model "families," or groups of models released by the same model provider. Models within a family are generally based on the same architecture and are often trained to behave similarly at the logit level. We ran experiments with models from the Llama model family, Llama-3.2-1B and Llama-3.2-3B,



(A) Experiments with a Llama model.

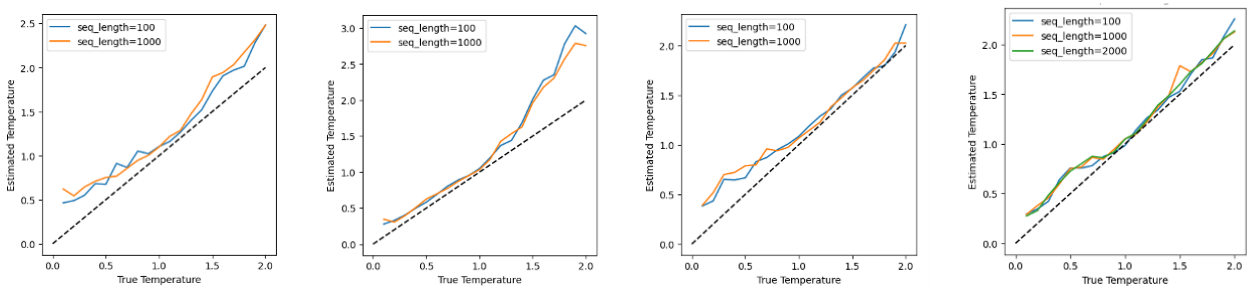


(B) Experiments with a Qwen model.

FIGURE 6.3. Experiments with language models mirroring the results from simulations in Figure 6.1. Top rows show true vs. estimated temperature at different temperatures. Bottom row shows theoretical and measured standard deviation at different temperatures across repeated experiments.

as well as from the Google's Gemma model family, gemma-2b and gemma-7b (Team et al., 2024). Experiments are set up similarly to the experiments in Section 6.6.1, where text is generated at varying lengths at temperatures from 0.1 to 2.0 and each temperature is used for 10 independent generations. Of note, due to memory limitations, some experiments don't have results for sequence lengths of 2,000. Figure 6.4 shows the results of these experiments. They indicate that models within the same family are able to reliably estimate the temperature of text generated by models within that same family.

Expanding our experiments to measure temperature across families tells a different story. In these tests, we used models from the Llama and Gemma families, as well as Mistral AI's Mistral family (Jiang et al., 2023). We specifically used Llama-3.2-1B from the Llama family, gemma-2b from the Gemma family, and Mistral-7B-v0.1 form the Mistral family. Again, the experimental setup is the same as in Section 6.6.1. Results from these experiments are in Figure 6.5. They show that temperature estimation is much less accurate across families than it is between families. Moreover, they show an inaccuracy pattern that seems to hold across experiments where temperature is over-estimated at low temperature and under-estimated at high temperatures. Specifically, true temperatures below about 0.75 tend to be estimated to be about 0.75, true temperatures above about 1.25 tend to be estimated to be about 1.25, and true temperatures between those tend to be relatively accurate, creating an S-shaped curve that is visible across experiments.



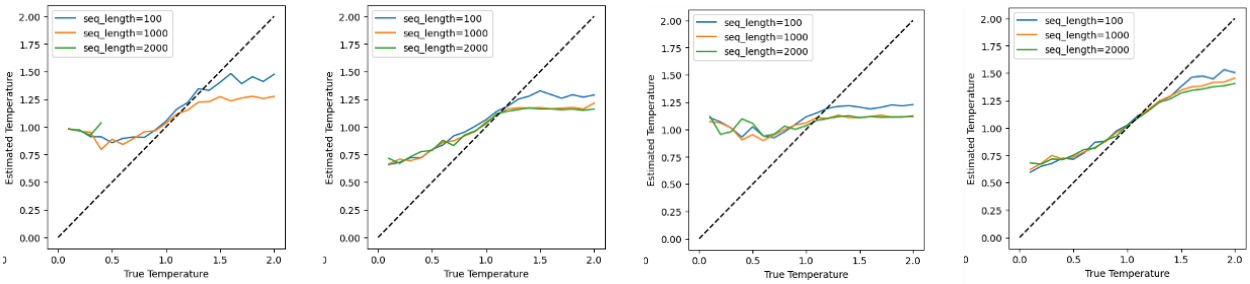
(A) Text generated by gemma-2b and temperature estimated by gemma-7b.

(B) Text generated by gemma-7b and temperature estimated by gemma-2b.

(C) Text generated by Llama-3.2-1B and temperature estimated by Llama-3.2-3B.

(D) Text generated by Llama-3.2-3B and temperature estimated by Llama-3.2-1B.

FIGURE 6.4. True vs. estimated temperature for temperature estimation experiments within Llama and Gemma model families. Within each model family, text is generated by one model and estimated by the other, and then the experiment is repeated in the other direction.



(A) Text generated by Llama-3.2-1B and temperature estimated by gemma-2b.

(B) Text generated by gemma-2b and temperature estimated by Llama-3.2-1B.

(C) Text generated by Llama-3.2-1B and temperature estimated by Mistral-7B-v0.1.

(D) Text generated by Mistral-7B-v0.1 and temperature estimated by Llama-3.2-1B.

FIGURE 6.5. True vs. estimated temperature for temperature estimation experiments across model families. Text is generated by a model in one model family and estimated by a model from another model family, and then the experiment is repeated in the other direction.

6.7. Discussion

Experiments indicate that our method of temperature estimation works well when the surrogate model used for estimation is either the same model or within the same model family as the model that generated the text. When the estimator uses a model outside of the family of the generating model, accuracy follows a consistent pattern: low true temperatures are overestimated,

high true temperatures are underestimated, and a middle band of temperatures is estimated relatively accurately, producing an S-shaped curve. Notably, this S-shaped pattern appears regardless of which of the two models in a cross-family pair is used as the generator and which is used as the surrogate.

One possible explanation for within-family success is that the generating model and the estimating model share similar logit distributions, both in terms of shape and scale. Models within a model family are often trained using a process called knowledge distillation, where a smaller "student" model is trained based on the knowledge of a larger "teacher" model, often through learning how to match the logit distribution of the larger model (Hinton et al., 2015). Since models within a family are trained to match a shared logit distribution, this could explain why they are able to be used to estimate each other's temperatures. This kind of logit matching doesn't exist across model families, so estimation in that setting is exposed to whatever mismatch exists between the two models' logit distributions.

This raises a separate question, however: why does a logit mismatch create the specific S-shaped pattern we observe? We see two plausible, non-exclusive explanations for this.

The first is a mismatch between models' confidence in what token comes next in a piece of text. If a surrogate model disagrees on the distribution of likelihoods for a next token compared to the generating model, then a "high confidence" next token and a "low confidence" according to the generator model won't look as low or high, respectively, to the surrogate model. To the surrogate model, all of the tokens are just as unaligned and just as likely/unlikely. To test this explanation for a given cross-family pair of models, one could compare the entropy each model assigns specifically to the tokens chosen by the other model during generation, in both directions, to compare the confidence each model assigns to next tokens.

The second possibility is that the S-curve reflects a more general property of MLE under model misspecification, rather than anything specific to the direction or source of a logit mismatch. Our theoretical guarantees in Section 6.3 assume the surrogate's logits are the generating model's true logits. When this assumption is violated, the resulting likelihood function is misspecified, and MLE under misspecification may not retain the property of being unbiased that we derived. One plausible consequence is that the estimator is pulled toward the value of τ that makes the observed

sequence appear most "unremarkable" under the surrogate's own raw logit landscape. For any surrogate, this would be a τ value near one, the setting that neither sharpens nor flattens its native distribution. If this explanation holds, one would expect the overall degree of mismatch between the two models' logit distributions (e.g., as measured by KL divergence) to correlate with the severity of the S-curve. In other words, more misaligned distributions would lead to a flatter S-curve, because a more highly misspecified MLE would more severely push towards "unremarkable" values.

The two explanations above address why a logit mismatch would produce this particular S-shaped pattern. A separate question is why the mismatch exists at all, which follow up work should also explore. One candidate is architectural difference: model providers each use their own architectural backbones, which could create differences in how tokens are processed and generated. Comparing models with the same architecture but different training strategies could help isolate this factor. Another candidate is tokenizer mismatch. Each model provider uses a distinct tokenizer, and our method relies on aligning the surrogate's predicted token at each step with the token actually chosen by the generating model. Differences in how text is segmented could introduce systematic misalignment between models with different tokenizers, independent of any difference in how the models represent language. Simple experiments mapping tokens between tokenizers in simulated settings could help isolate this effect from genuine representational mismatch.

If these more mechanistic explanations fail to fully account for the cross-family gap, a more speculative possibility is that the gap reflects a fundamental mismatch in how different models represent language, potentially indicating that language models understand language in at least partially incompatible ways. There is already evidence that LLMs display a preference for their own outputs (Ye et al., 2025; Wataoka et al., 2025), and a deeper representational mismatch of this kind could offer one explanation for that preference. Experiments assessing LLMs' understanding of text generated by other models could help validate that hypothesis. However, the more immediate, mechanistic explanations above should be the first step for future work.

It is important to note some of the assumptions and limitations of the experiments conducted in this work. One major assumption is the decoding strategy used in generating text. The decoding strategy used was a sampling strategy that didn't use any of the methods commonly found in modern language models, including top-k and top-p filtering. Moreover, prompts are mostly ignored

in this work. Language models are almost exclusively used with some kind of prompt, and follow up work could explore the effect of prompting on the accuracy of this temperature estimation method. This work also only explores temperature estimation on freeform text. LLMs frequently generate more structured text, like code.

Even with these limitations, this work introduces a method for estimating the temperature of LLM-generated text that doesn't require the training of an external classifier and is theoretically grounded. It is a step towards a generalizable method for temperature estimation, an important part of characterizing LLM-generated outputs, and also potentially opens up avenues for improving our understanding of how LLMs understand each other.

APPENDIX A

Appendix for Chapter 2

Table A.1 shows all of the electronic health record features that were used during model training.

Figure A.1 shows the same plot as Figure 2.3 but for all modality combinations. Further, Figure A.2 shows Brier scores for all modality combinations. Of note, calibration as measured by Brier score improves from uni- to bi- to trimodal models.

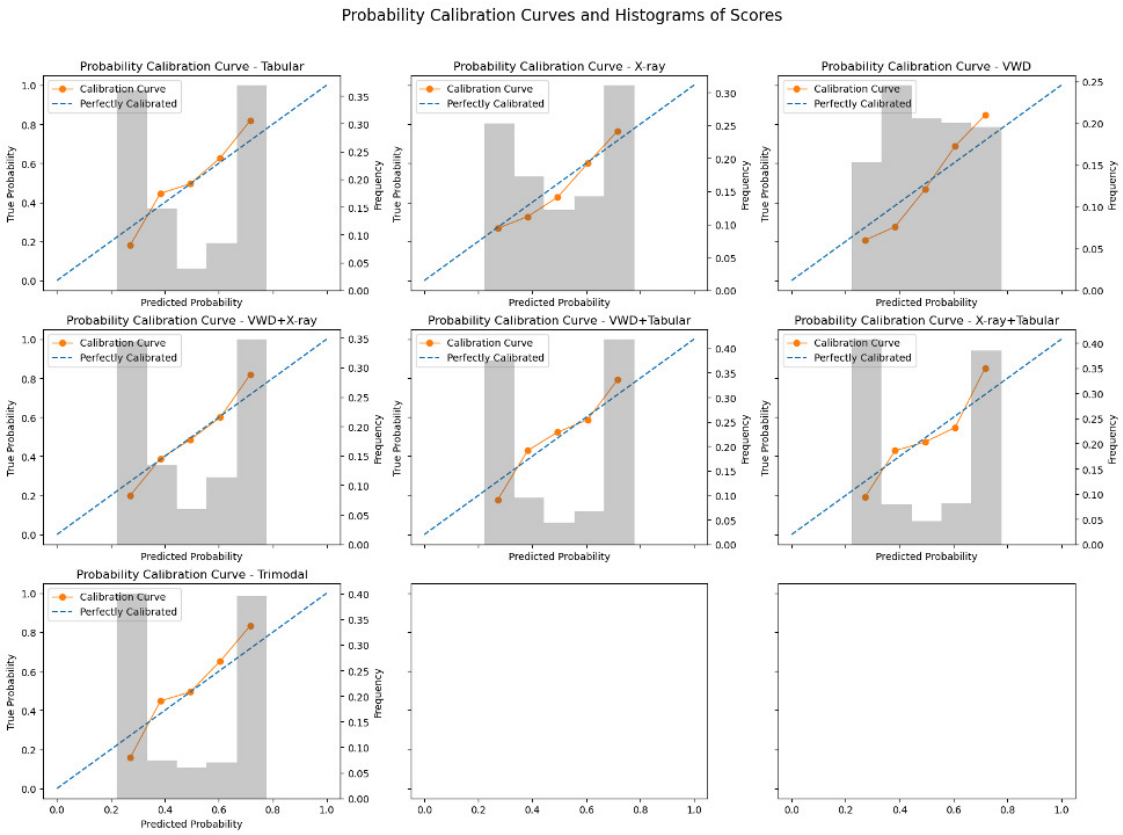


FIGURE A.1. Calibration plots for models trained on all modality combinations.

Numerical Features	Categorical Features
total_intake_ml	smoking_usage
total_output_ml	tobacco_user
sf	age_at_admsn_yr
sf97	sex
osi	race
osi97	ethnicity
temperature_celsius	adi_state_score
respiratory_rate	ca_hpi_percentile
heart_rate	ca_hpi_quartile
fio2	elixhauser_weighted_score
spo2	elixhauser_comorbidity_count
lab_pf_ratio_res	
lab_white_blood_cell_res	
lab_neutrophil_abs_manual_res	
lab_neutrophil_abs_auto_res	
lab_bands_percent_res	
lab_hemoglobin_res	
lab_platelet_res	
lab_inr_res	
lab_sodium_res	
lab_potassium_res	
lab_chloride_res	
lab_bicarbonate_res	
lab_e_gfr_res	
lab_serum_glucose_res	
lab_bilirubin_total_res	
lab_asparate_transminase_res	
lab_alanine_transferase_res	
lab_alkaline_phosphatase_res	
lab_lipase_res	
lab_bnp_res	
lab_creatinine_res	
lab_bun_res	
bp_map_cuff	
bp_cuff_systolic	
bp_cuff_diastolic	
bp_map_a_line	
bp_a_line_systolic	
bp_a_line_diastolic	

TABLE A.1. Full list of numerical and categorical features in EHR data.

Modality	Brier Score
VWD	0.2098
Tabular	0.1807
X-ray	0.2113
X-ray+Tabular	0.1760
VWD+Tabular	0.1816
VWD+X-ray	0.1895
Trimodal	0.1696

TABLE A.2. Brier score for each modality combination.

APPENDIX B

Appendix for Chapter 3

B.1. Proofs

B.1.1. Proof of Proposition 3.5.2.

Proof. By Markov's inequality,

$$\mathbb{P}_{X\sim\mu_0}\big(\|g(f(X))-X\|>\varepsilon\big)\le\frac{\mathbb{E}\|g(f(X))-X\|^2}{\varepsilon^2}, \tag{B.1}$$

$$\mathbb{P}_{Y\sim\mu_1}\big(\|f(g(Y))-Y\|>\varepsilon\big)\le\frac{\mathbb{E}\|f(g(Y))-Y\|^2}{\varepsilon^2}. \tag{B.2}$$

Summing these inequalities yields:

$$\mathbb{P}_{X\sim\mu_0}\big(\|g(f(X))-X\|>\varepsilon\big)+\mathbb{P}_{Y\sim\mu_1}\big(\|f(g(Y))-Y\|>\varepsilon\big)\le\frac{\mathcal{C}(f,g)}{\varepsilon^2}\le\frac{\delta^2}{\varepsilon^2}, \tag{B.3}$$

which completes the proof. □

B.1.2. Proof of Lemma 3.5.2.1.

Proof. By the source-side expansion,

$$g^+(f^+(X))-X=R_0(X)-\alpha\widetilde{W}_0(X)+\alpha^2\widetilde{Q}_0(X)+\mathcal{R}_0^\alpha(X),$$

where $\|\mathcal{R}_0^\alpha\|_{L^2(\mu_0)}\le C\alpha^3$. Using

$$\|a-b\|^2\ge\frac{1}{2}\|b\|^2-\|a\|^2$$

with

$$a=R_0+\alpha^2\widetilde{Q}_0+\mathcal{R}_0^\alpha,\qquad b=\alpha\widetilde{W}_0,$$

we obtain

$$\mathbb{E}\|g^+(f^+(X))-X\|^2\ge\frac{1}{2}\alpha^2\mathbb{E}\|\widetilde{W}_0(X)\|^2-\mathbb{E}\|R_0+\alpha^2\widetilde{Q}_0+\mathcal{R}_0^\alpha\|^2.$$

Since $\mathbb{E}\|R_0\|^2 \le \mathcal{C}(z) \le \delta^2$ and $\mathbb{E}\|\widetilde{Q}_0\|^2 \le C_Q$,

$$\mathbb{E}\|R_0 + \alpha^2\widetilde{Q}_0 + \mathcal{R}_0^\alpha\|^2 \le \mathbb{E}\|R_0\|^2 + C\alpha^2\delta + C\alpha^3.$$

Thus

$$\mathbb{E}\|g^+(f^+(X)) - X\|^2 \ge \frac{1}{2}\alpha^2\mathbb{E}\|\widetilde{W}_0(X)\|^2 - \mathbb{E}\|R_0(X)\|^2 - C\alpha^2\delta - C\alpha^3.$$

The same argument on the target side gives

$$\mathbb{E}\|f^+(g^+(Y)) - Y\|^2 \ge \frac{1}{2}\alpha^2\mathbb{E}\|\widetilde{W}_1(Y)\|^2 - \mathbb{E}\|R_1(Y)\|^2 - C\alpha^2\delta - C\alpha^3.$$

Adding the two inequalities yields

$$\mathcal{C}(z^+) \ge \frac{1}{2}\alpha^2\widetilde{E}_{\text{mismatch}}(z) - \mathcal{C}(z) - C\alpha^2\delta - C\alpha^3.$$

Using $\mathcal{C}(z) \le \delta^2$ and $\mathcal{C}(z^+) \le \delta^2 + \varepsilon$, we obtain

$$\frac{1}{2}\alpha^2\widetilde{E}_{\text{mismatch}}(z) \le 2\delta^2 + \varepsilon + C\alpha^2\delta + C\alpha^3.$$

Dividing by α^2 proves the result. □

B.1.3. Cyclic Monotonicity Characterization.

THEOREM B.1.1 (Cyclic monotonicity and quadratic OT). *Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and assume $\mu_0 \ll \mathcal{L}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable map satisfying*

$$f_\#\mu_0 = \mu_1.$$

Then f is the unique quadratic-cost optimal transport map from μ_0 to μ_1 if and only if f is cyclically monotone μ_0-almost surely. Equivalently, $f = \nabla\varphi$ μ_0-almost surely for some convex potential φ, and f minimizes

$$\inf_{T_\#\mu_0=\mu_1} \int \frac{1}{2}\|T(x) - x\|^2 \, d\mu_0(x).$$

B.2. Training Algorithm

Algorithm 2 Synchronized Forward–Backward Optimal Transport

Require: distributions μ_0, μ_1; forward map network $f_\theta : \mathcal{X} \to \mathcal{X}$; backward map network $g_\phi : \mathcal{X} \to \mathcal{X}$; discriminators network $D^0_{\omega_0}, D^1_{\omega_1} : \mathcal{X} \to (0,1)$; Training iterations K; Initial weight λ_0

1: **for** $k = 1, \ldots, K$ **do**

2: Sample batches $\{x_0^i\}_{i=1}^B \sim \mu_0$ and $\{x_1^i\}_{i=1}^B \sim \mu_1$.

3: Compute discriminator losses

$$\mathcal{L}_{D^1_{\omega_1}} \leftarrow -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log D^1_{\omega_1}(x_1^i) + \log\big(1 - D^1_{\omega_1}(f_\theta(x_0^i))\big)\Big],$$

$$\mathcal{L}_{D^0_{\omega_0}} \leftarrow -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log D^0_{\omega_0}(x_0^i) + \log\big(1 - D^0_{\omega_0}(g_\phi(x_1^i))\big)\Big],$$

4: Update $D^1_{\omega_1}, D^0_{\omega_0}$ by $\frac{\partial \mathcal{L}_{D^1_{\omega_1}}}{\partial \omega_1}$, $\frac{\partial \mathcal{L}_{D^0_{\omega_0}}}{\partial \omega_0}$ respectively.

5: Compute adversarial generator loss

$$\mathcal{L}_{\text{adv}} = -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log D^1_{\omega_1}(f_\theta(x_0^i)) + \log D^0_{\omega_0}(g_\phi(x_1^i))\Big].$$

6: Compute consistency loss

$$\mathcal{L}_{\text{cyc}} = \frac{1}{2B}\sum_{i=1}^{B}\Big[\|g_\phi(f_\theta(x_0^i)) - x_0^i\|_2^2 + \|f_\theta(g_\phi(x_1^i)) - x_1^i\|_2^2\Big].$$

7: Compute transport cost regularization

$$\mathcal{L}_{\text{cost}} = \frac{1}{2B}\sum_{i=1}^{B}\Big[\|f_\theta(x_0^i) - x_0^i\|_2^2 + \|g_\phi(x_1^i) - x_1^i\|_2^2\Big].$$

8: Update f_θ and g_ϕ by minimizing

$$\mathcal{L}_{\text{map}} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{cyc}} + {\lambda_0}^k \mathcal{L}_{\text{cost}}.$$

9: **end for**

10:

11: **return** forward and backward generator f_θ, g_ϕ.

B.3. Additional Experimental Results

B.3.1. Pneumonia. Quantitative results in 3.6.7 only included the transport from healthy to unhealthy images. Table B.1 includes results in both directions, where *Pos* → *Neg* indicates transporting from healthy to unhealthy, and *Neg* → *Pos* indicates transporting from unhealthy to healthy. Baseline predictions for true images in each class can be found in Table 3.6.

Method	Direction	Cost ↓	|Enrich-k100 − 0.5| ↓	TPS ↑
ICNN OT	Pos → Neg	73142.1836	0.4986	0.1517 ± 0.0260
	Neg → Pos	65095.5671	0.4984	0.1513 ± 0.0409
Neural OT	Pos → Neg	1623.5243	0.0507	0.4180 ± 0.1811
	Neg → Pos	1984.1612	0.3246	0.4762 ± 0.1860
Ours	Pos → Neg	1472.9777	0.0133	0.4484 ± 0.1856
	Neg → Pos	1717.2522	0.3169	0.4098 ± 0.1873

TABLE B.1. Quantitative comparison between ICNN OT, Neural OT, and our method on the chest x-ray healthy to pneumonia transition task, broken down by transport direction.

APPENDIX C

Appendix for Chapter 5

C.1. Proof of Proposition 5.3.1

PROOF. To start, define the Bayes error as

$$P_e := \mathbb{E}_{X_{\text{MarI}}}\Big[\min\big\{P(Z=0 \mid X_{\text{MarI}}),\, P(Z=1 \mid X_{\text{MarI}})\big\}\Big] = 1 - P_{acc}.$$

In addition, for each x, let $p(x) := P(Z=1 \mid X_{\text{MarI}} = x) \in [0,1]$ be the conditional probability of $\{Z=1\}$ given $\{X_{\text{MarI}} = x\}$. Then it follows from Z being binary that $H(Z \mid X_{\text{MarI}} = x) = H_2\big(p(x)\big)$. Denote $m(X_{\text{MarI}}) := \min\big\{P(Z=0 \mid X_{\text{MarI}}),\, P(Z=1 \mid X_{\text{MarI}})\big\}$. Since H_2 is concave, it follows from Jensen's inequality that

$$\begin{aligned} H(Z \mid X_{\text{MarI}}) &= \mathbb{E}_{X_{\text{MarI}}}\big[H_2(p(X_{\text{MarI}}))\big] \\ &= \mathbb{E}_{X_{\text{MarI}}}\big[H_2(m(X_{\text{MarI}}))\big] \\ &\le H_2\big(\mathbb{E}_{X_{\text{MarI}}}[m(X_{\text{MarI}})]\big) \\ &= H_2(P_e). \end{aligned}$$

where the second equality holds due to the fact that $H_2(p) = H_2(1-p)$. Now, since $I(X_{\text{MarI}}; Z) = H(Z) - H(Z \mid X_{\text{MarI}})$ and $H(Z) = H_2(\pi)$, we obtain

$$H_2(\pi) - I(X_{\text{MarI}}; Z) = H(Z \mid X_{\text{MarI}}) \;\le\; H_2(P_e).$$

Since $P_e \in [0, \frac{1}{2}]$ and H_2 is strictly increasing on this interval, by applying the inverse H_2^{-1}, we have

$$P_e \;\ge\; H_2^{-1}\big(H_2(\pi) - I(X_{\text{MarI}}; Z)\big).$$

Finally, by $P_{acc} = 1 - P_e$, we have

$$P_{acc} \leq 1 - H_2^{-1}\big(H_2(\pi) - I(X_{\mathrm{MarI}}; Z)\big).$$

This proves the stated inequality. The particular case $\pi = \frac{1}{2}$ follows from $H_2(\frac{1}{2}) = 1$.

It remains to show that the upper bound is tight. Indeed, fix an arbitrary $I \in [0, H_2(\pi)]$. Choose $p^\star \in \left[\frac{1}{2}, 1\right]$ such that $H_2(p^\star) = H_2(\pi) - I$. Construct $P_{Z|X_{\mathrm{MarI}}}$ such that $P(Z = 1 \mid X_{\mathrm{MarI}}) \in \{p^\star, 1-p^\star\}$ with probabilities chosen to match the prior π. Then $H(Z \mid X_{\mathrm{MarI}}) = H_2(p^\star)$ and $I(X_{\mathrm{MarI}}; Z) = I$, while the Bayes error satisfies $P_e = \min\{p^\star, 1-p^\star\} = H_2^{-1}(H(Z) - I)$. Hence, equality holds in the bound. □

C.2. Generalization of Theorem 5.3.2 to unlearn set neighborhood

Here, we show that the self-perplexity gap guarantee provided by Theorem 5.3.2 can be generalized to a neighborhood of u rather than u itself:

THEOREM C.2.1 (MarI controls neighborhood-perplexity gap). *Draw* $U := \{U_t\}_{t=1}^T$ *with* $U_t \sim p_t^u$ *independently across* $t \in [T]$ *and suppose* $\max_{t,x} \left\{\frac{p_t^u(x)}{p_t^r(x)} \vee \frac{p_t^r(x)}{p_t^u(x)}\right\} =: M < \infty$. *Let* $C := \max_{t,x:\, p_t^u(x)>0} \left[\log \frac{p_t^r(x)}{p_t^u(x)}\right]^2 < \infty$. *Then, for any* $\epsilon > 0$, *with probability at least* $1 - 2\exp\big(-T\epsilon^2/(2C)\big)$,

$$\big|S_\theta(U, u) - S_\theta(U, r)\big| \;\leq\; \Big(\log M\Big) \frac{M}{M-1} \frac{\sqrt{2}}{1-\alpha} \sqrt{I(X_{\mathrm{MarI}}; Z)} \;+\; \epsilon.$$

We start with the following three lemmata that are needed for the proof of Theorem C.2.1:

LEMMA C.2.1.1 (Point-wise KL bound). *Let* p, q *be two probability distributions over a finite set* V *such that* $\frac{p(x)}{q(x)} \in [1, M]$ *for every* $x \in V$ *for some constant* $M > 1$. *Then for every* $x \in V$

$$p(x) \log \frac{p(x)}{q(x)} \;\leq\; (\log M) \frac{M}{M-1} \big[p(x) - q(x)\big]. \tag{C.1}$$

PROOF. Fix $x \in V$ and set $y := \dfrac{p(x)}{q(x)} \in [1, M]$. Inequality (C.1) is equivalent to

$$y \log y \;\leq\; \frac{M}{M-1} (\log M)\,(y-1), \qquad \forall\, y \in [1, M]. \tag{C.1}$$

For $y > 1$ let $g(y) := \frac{y \log y}{y-1}$ and set $g(1) := \lim_{y \to 1^+} g(y) = 1$. We show that g is strictly increasing on $[1, M]$. Indeed, compute $g'(y) = \frac{(y-1)-\log y}{(y-1)^2}$. Since $\log y < y - 1$ for all $y > 1$, we have $g'(y) > 0$; thus, g is strictly increasing. Because g is increasing and $y \in [1, M]$, we have

$$g(y) \;\le\; g(M) = \frac{M \log M}{M-1}.$$

Multiplying both sides by $y-1$ yields (C.1), which is precisely (C.1) after reinstating $y = p(x)/q(x)$. Therefore, (C.1) holds for every $x \in V$. This completes the proof. □

LEMMA C.2.1.2. **(Total Variation is controlled by Jensen-Shannon Divergence)** *For any two probability measures p, q on a finite set, we have*

$$\|p - q\|_{TV} \;\le\; \sqrt{2 \text{ JSD}(p, q)},$$

where $\text{JSD}(p, q) := \frac{1}{2} \text{D}_{\text{KL}}(p \| m) + \frac{1}{2} \text{D}_{\text{KL}}(q \| m)$, $m := \frac{p+q}{2}$ *and* $\text{D}_{\text{KL}}(p \| q) := \sum_v p(v) \log \frac{p(v)}{q(v)}$, *denotes the Jensen–Shannon divergence.*

PROOF. Let $m = \frac{p+q}{2}$. Pinsker's inequality gives $\|p - m\|_1^2 \le 2 \text{ D}_{\text{KL}}(p \| m)$ and analogously for q. Hence

$$\text{JSD}(p, q) \;\ge\; \tfrac{1}{4}\Big[\|p - m\|_1^2 + \|q - m\|_1^2\Big] \;=\; \tfrac{1}{8} \|p - q\|_1^2,$$

because $p - m = \frac{p-q}{2}$ and $q - m = -\frac{p-q}{2}$. Since $\|p - q\|_{TV} = \frac{1}{2}\|p - q\|_1$, it follows that $\|p - q\|_{TV}^2 \le 2 \text{ JSD}(p, q)$. □

LEMMA C.2.1.3 (Exact TV scaling under mixture). *If $p^d = \alpha p^r + (1 - \alpha) p^u$ with $\alpha \in (0, 1)$, then*

$$\|p^u - p^r\|_{TV} \;=\; \frac{1}{1 - \alpha} \|p^d - p^r\|_{TV}.$$

PROOF. $p^d - p^r = (1 - \alpha)(p^u - p^r)$. Taking ℓ_1-norms and dividing by 2 yields the identity. □

Now, we are ready to prove Theorem C.2.1:

PROOF. Define $Y_t := \log \frac{p_t^r(U_t)}{p_t^u(U_t)}$, so that

$$S_\theta(U, u) - S_\theta(U, r) \;=\; \frac{1}{T} \sum_{t=1}^{T} Y_t.$$

Since $U_t \sim p_t^u$, $\mathbb{E}[Y_t] = \sum_x p_t^u(x) \log \frac{p_t^r(x)}{p_t^u(x)} = -\operatorname{D_{KL}}(p_t^u \| p_t^r)$, hence

$$\mathbb{E}\big[S_\theta(U,u) - S_\theta(U,r)\big] = -\frac{1}{T}\sum_{t=1}^{T} \operatorname{D_{KL}}(p_t^u \| p_t^r).$$

Now, by the assumption $\max_{t,x} \max\left\{\frac{p_t^u(x)}{p_t^r(x)}, \frac{p_t^r(x)}{p_t^u(x)}\right\} \le M$, we have $p_t^r(x) > 0$ for $p_t^u(x)$-a.e. x for all t. Therefore, for all t, we have $\log \frac{p_t^r(x)}{p_t^u(x)} < \infty$ and taking the maximum over $t \in [T]$, we obtain $C := \max_{t,\, x:\, p_t^u(x)>0} \left[\log \frac{p_t^r(x)}{p_t^u(x)}\right]^2 < \infty$. It then follows from the definition of Y_t that $|Y_t| \le \sqrt{C}$ a.s.. Hoeffding's inequality for independent bounded variables yields, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Y_t - \mathbb{E}\frac{1}{T}\sum_{t=1}^{T} Y_t\right| \ge \epsilon\right) \le 2\exp\left(-\frac{T\,\epsilon^2}{2C}\right).$$

Using $\big||a| - b\big| \le |a-b|$ for $b \ge 0$, we have

$$\big|S_\theta(U,u) - S_\theta(U,r)\big| \le \frac{1}{T}\sum_{t=1}^{T} \operatorname{D_{KL}}(p_t^u \| p_t^r) + \epsilon.$$

with probability at least $1 - 2\exp(-\frac{T\,\epsilon^2}{2C})$.

Now, for each t, let $A_t = \{x : p_t^u(x) \ge p_t^r(x)\}$. Then by Lemma C.2.1.1, we have

$$\operatorname{D_{KL}}(p_t^u \| p_t^r) \le \kappa(M) \sum_{x \in A_t} \big(p_t^u(x) - p_t^r(x)\big) \le \kappa(M)\, \|p_t^u - p_t^r\|_{TV}.$$

Averaging in t gives

$$\frac{1}{T}\sum_{t=1}^{T} \operatorname{D_{KL}}(p_t^u \| p_t^r) \le \kappa(M)\, \frac{1}{T}\sum_{t=1}^{T} \|p_t^u - p_t^r\|_{TV}.$$

Finally, it follows from Lemma C.2.1.3 and Lemma C.2.1.2 that

$$\frac{1}{T}\sum_{t=1}^{T} \|p_t^u - p_t^r\|_{TV} = \frac{1}{1-\alpha}\,\frac{1}{T}\sum_{t=1}^{T} \|p_t^d - p_t^r\|_{TV} \le \frac{\sqrt{2}}{1-\alpha}\,\frac{1}{T}\sum_{t=1}^{T} \sqrt{\operatorname{JSD}(p_t^d, p_t^r)}.$$

By Jensen's inequality, $\frac{1}{T}\sum_t \sqrt{\operatorname{JSD}(p_t^d, p_t^r)} \le \sqrt{\frac{1}{T}\sum_t \operatorname{JSD}(p_t^d, p_t^r)}$. Combining the displays proves the claim with $I(X_{\mathrm{MarI}}; Z) = \frac{1}{T}\sum_t \operatorname{JSD}(p_t^d, p_t^r)$. □

C.3. Proof of Theorem 5.3.2

Here, we provide the proof for Theorem 5.3.2:

PROOF. By the mean value theorem, for each t there exists $\xi_t \in [\min\{p_t^u(u_t), p_t^r(u_t)\}, 1] \subseteq [\gamma, 1]$ such that

$$\begin{aligned}\left|\log p_t^u(u_t) - \log p_t^r(u_t)\right| &= \frac{\left|p_t^u(u_t) - p_t^r(u_t)\right|}{\xi_t} \\ &\le \frac{\left|p_t^u(u_t) - p_t^r(u_t)\right|}{\gamma} \\ &\le \frac{\|p_t^u - p_t^r\|_1}{\gamma} \\ &= \frac{2\,\|p_t^u - p_t^r\|_{TV}}{\gamma}.\end{aligned}$$

Averaging over t,

$$\left|S_\theta(u,u) - S_\theta(u,r)\right| \;\le\; \frac{2}{\gamma}\,\frac{1}{T}\sum_{t=1}^{T}\|p_t^u - p_t^r\|_{TV}.$$

Apply Lemma C.2.1.3 followed by Lemma C.2.1.2 and Jensen's inequality:

$$\frac{1}{T}\sum_t \|p_t^u - p_t^r\|_{TV} \;=\; \frac{1}{1-\alpha}\,\frac{1}{T}\sum_t \|p_t^d - p_t^r\|_{TV} \;\le\; \frac{\sqrt{2}}{1-\alpha}\sqrt{\frac{1}{T}\sum_t \mathrm{JSD}(p_t^d, p_t^r)}.$$

Combining yields the claim. □

C.4. Theoretical error bound between position-wise vs. pooled MarI

Here, we provide the theoretical analysis of the error between MarI and pooled MarI. The following result shows that, under a mild assumption, the error of using the pooled MarI to estimate MarI is bounded by the sequence-wise density variance:

THEOREM C.4.1 (Pooling error bound). *For each t, set $m_t := \frac{1}{2}(p_t^d + p_t^r)$ and $\overline{m} := \frac{1}{2}(\overline{p}^{\,d} + \overline{p}^{\,r})$, where $\overline{p}^{\,d} := \frac{1}{T}\sum_{t=1}^{T} p_t^d$ and $\overline{p}^{\,r} := \frac{1}{T}\sum_{t=1}^{T} p_t^r$. Assume the uniform overlap condition*

(C.2)

$$\beta \;:=\; \min\left\{ \inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\,m_t(x) + \lambda\,\overline{m}(x)\big],\right.$$

(C.3)

$$\left. \inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\,p_t^d(x) + \lambda\,\overline{p}^{\,d}(x)\big],\; \inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\,p_t^r(x) + \lambda\,\overline{p}^{\,r}(x)\big] \right\} \;>\; 0.$$

Define the (averaged) ℓ_2-deviation terms

$$V_d \;:=\; \frac{1}{T}\sum_{t=1}^{T}\big\|p_t^{\,d}-\overline{p}^{\,d}\big\|_2^2, \qquad V_r \;:=\; \frac{1}{T}\sum_{t=1}^{T}\big\|p_t^{\,r}-\overline{p}^{\,r}\big\|_2^2.$$

Then

$$0 \;\le\; I(X_{\mathrm{MarI}};Z) \;-\; I(\overline{X}_{\mathrm{MarI}};Z) \;\le\; \frac{1}{4\,\beta}\Big(V_d+V_r\Big), \tag{C.4}$$

where $I(X_{\mathrm{MarI}};Z)=\frac{1}{T}\sum_{t=1}^{T}\mathrm{JSD}(p_t^{\,d},p_t^{\,r})$ *and* $I(\overline{X}_{\mathrm{MarI}};Z)=\mathrm{JSD}(\overline{p}^{\,d},\overline{p}^{\,r})$.

Proof. The lower bound $I(\overline{X}_{\mathrm{MarI}};Z)\le I(X_{\mathrm{MarI}};Z)$ follows directly from the data-processing inequality. For the upper bound, write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathrm{JSD}(p_t^{\,d},p_t^{\,r}) \;-\; \mathrm{JSD}(\overline{p}^{\,d},\overline{p}^{\,r})\\
=&\frac{1}{T}\sum_{t=1}^{T}\Big[\underbrace{H(m_t)-H(\overline{m})}_{(A_t)}-\frac{1}{2}\underbrace{\big(H(p_t^{\,d})-H(\overline{p}^{\,d})\big)}_{(B_t)}-\frac{1}{2}\underbrace{\big(H(p_t^{\,r})-H(\overline{p}^{\,r})\big)}_{(C_t)}\Big].
\end{aligned}$$

Let $\Delta_m := m_t-\overline{m}$, $\Delta_d := p_t^{\,d}-\overline{p}^{\,d}$, $\Delta_r := p_t^{\,r}-\overline{p}^{\,r}$. Since H is twice-differentiable, the second-order Taylor expansion around the pooled densities yields

$$H(a) \;=\; H(a')+\langle\nabla H(a'),a-a'\rangle+\tfrac{1}{2}(a-a')^{\top}\nabla^2H(s)(a-a'),$$

for some s on the line segment between a' and a. Note that H is concave, $\nabla^2H(s)$ is negative semidefinite and diagonal with entries $-1/s(x)$. By the overlap assumption (C.2), every coordinate along the segments between m_t and $\overline{m}$ and between $p_t^{\,s}$ and $\overline{p}^{\,s}$ ($s\in\{d,r\}$) is at least β, hence

$$-\,(a-a')^{\top}\nabla^2H(s)(a-a') \;\le\; \frac{1}{\beta}\sum_{x\in V}\big(a(x)-a'(x)\big)^2 \;\le\; \frac{1}{\beta}\,\|a-a'\|_2^2,$$

and therefore the (negative) Taylor remainders satisfy

$$H(a)-H(a')-\langle\nabla H(a'),a-a'\rangle \;\ge\; -\frac{1}{2\beta}\,\|a-a'\|_2^2. \tag{C.5}$$

Applying (C.5) to the three entropy differences and averaging in t, the first-order terms vanish because $\frac{1}{T}\sum_t \Delta_d = 0$, $\frac{1}{T}\sum_t \Delta_r = 0$, and $\frac{1}{T}\sum_t \Delta_m = 0$. Thus,

$$\begin{aligned}
\frac{1}{T}\sum_t \big[(A_t) - (B_t)/2 - (C_t)/2\big] \leq & \frac{1}{T}\sum_t \frac{1}{2}\Big(-\mathcal{R}_d(t)\Big) \;+\; \frac{1}{T}\sum_t \frac{1}{2}\Big(-\mathcal{R}_r(t)\Big) - \frac{1}{T}\sum_t \mathcal{R}_m(t) \\
\leq & -\frac{1}{2}\Big(\frac{1}{T}\sum_t \Big(\mathcal{R}_d(t)\Big) \;+\; \frac{1}{T}\sum_t \Big(\mathcal{R}_r(t)\Big)\Big) \\
\leq & \frac{1}{4\beta}\Big(\frac{1}{T}\sum_t \Big(||p_t^d - \overline{p}^d||_2^2\Big) \;+\; \frac{1}{T}\sum_t \Big(||p_t^r - \overline{p}^r||_2^2\Big)\Big) \\
= & \frac{1}{4\beta}\,(V_d + V_r).
\end{aligned}$$

where $\mathcal{R}_s(t) := H(p_t^s) - H(\overline{p}^s) - \langle \nabla H(\overline{p}^s), p_t^s - \overline{p}^s\rangle \leq 0$ for $s \in \{d, r\}$ and $\mathcal{R}_m(t) := H(m_t) - H(\overline{m}) - \langle \nabla H(\overline{m}), m_t - \overline{m}\rangle \leq 0$. □

C.5. Theoretical Guarantee provided by pooled MarI

In general, there is *no non-trivial (distribution–free)* theoretical guarantee for the sequence–level perplexity gaps can be stated solely in terms of the pooled MarI $I(\overline{X}_{\mathrm{MarI}}; Z) = \mathrm{JSD}(\overline{p}^d, \overline{p}^r)$ without any additional assumption, such as the bounded ℓ^2-deviation in Theorem C.4.1 above.

Nonetheless, the pooled MarI *can* certify the following *word–level forgetting* (for particular tokens) forgetting:

THEOREM C.5.1 (Word-level provable unlearning via pooled MarI). *Fix a token $w \in V$ and assume $\overline{p}^u(w) \wedge \overline{p}^r(w) =: \overline{\gamma}_w \in (0, 1]$. Then*

$$\Big|\, \log \overline{p}^u(w) - \log \overline{p}^r(w) \,\Big| \;\leq\; \frac{2}{\overline{\gamma}_w\,(1-\alpha)}\,\sqrt{2\,I(\overline{X}_{\mathrm{MarI}}; Z)}\,.$$

PROOF. It follows from mean value theorem for $x \mapsto \log x$ on $[\overline{\gamma}_w, 1]$ that

$$|\log \overline{p}^u(w) - \log \overline{p}^r(w)| \leq |\overline{p}^u(w) - \overline{p}^r(w)|/\overline{\gamma}_w.$$

Since $\overline{p}^u - \overline{p}^r = (\overline{p}^d - \overline{p}^r)/(1-\alpha)$, we have

$$|\overline{p}^u(w) - \overline{p}^r(w)| \;\leq\; \|\overline{p}^u - \overline{p}^r\|_1 = \frac{1}{1-\alpha}\,\|\overline{p}^d - \overline{p}^r\|_1 = \frac{2}{1-\alpha}\,\|\overline{p}^d - \overline{p}^r\|_{TV}.$$

Finally, combining the above two inequalities, we have

$$|\log \overline{p}^{\,u}(w) - \log \overline{p}^{\,r}(w)| \leq \frac{2}{\overline{\gamma}_w(1-\alpha)} \, \|\overline{p}^{\,d} - \overline{p}^{\,r}\|_{TV} \leq \frac{2}{\overline{\gamma}_w(1-\alpha)} \sqrt{2\, I(\overline{X}_{\mathrm{MarI}}; Z)},$$

where the last inequality follows from the fact that

$$\|\overline{p}^{\,d} - \overline{p}^{\,r}\|_{TV} \leq \sqrt{2 \ \mathrm{JSD}(\overline{p}^{\,d}, \overline{p}^{\,r})} = \sqrt{2\, I(\overline{X}_{\mathrm{MarI}}; Z)}.$$

□

APPENDIX D

Appendix for Chapter 6

D.1. Deriving An Equation for Fisher Information

Taking the log of the likelihood yields:

$$\log P(X|\tau) = \sum_s \log p_{x_s,s}(\tau)$$

And for each step s this gives:

$$\log p_{x_s,s}(\tau) = \frac{l_{x_s,s}}{\tau} - \log\left(\sum_{j=1}^{K} e^{l_{js}/\tau}\right)$$

We'll drop the s subscript to solve for a single logit in the sequence.

The first derivative is:

$$\frac{\partial}{\partial\tau}\log p_{x_s}(\tau) = \frac{1}{\tau^2}\left(-l_{x_s} + \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}\right)$$

Full work for the derivative in Section D.2.

Simplifying by defining a new term:

$$W(\tau) = \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}$$

For the whole sequence, the log-likelihood gradient is:

$$\frac{\partial}{\partial \tau} \log P(X|\tau) = \sum_s \frac{1}{\tau^2}(-l_{x_s,s} + W_s(\tau))$$

So the Fisher information is:

$$I(\tau) = \mathbb{E}\left[\left(\sum_s \frac{1}{\tau^2}(-l_{x_s,s} + W_s(\tau))\right)^2\right] = \frac{1}{\tau^4}\mathbb{E}\left[\left(\sum_s (-l_{x_s,s} + W_s(\tau))\right)^2\right]$$

Assuming each step is independent, this then gives:

$$I(\tau) = \frac{1}{\tau^4}\sum_s \mathbb{E}\left[(-l_{x_s,s} + W_s(\tau))^2\right]$$

And for each step this then gives:

$$I(\tau) = \frac{1}{\tau^4}\mathbb{E}\left[(-l_{x_s,s} + W_{j,s})^2\right] = \frac{1}{\tau^4}\sum_{i=1}^{K} p_{i,s}(\tau)(-l_{i,s} + W_s(\tau))^2$$

D.2. Full Derivative Calculations

Starting with a single term of the log-likelihood:

$$\log p_i(\tau) = \frac{l_i}{\tau} - \log\left(\sum_{j=1}^{K} e^{l_j/\tau}\right).$$

The first derivative with respect to τ has two terms:

$$\frac{\partial}{\partial \tau}\log p_i(\tau) = \frac{\partial}{\partial \tau}\left(\frac{l_i}{\tau}\right) - \frac{\partial}{\partial \tau}\log\left(\sum_{j=1}^{K} e^{l_j/\tau}\right).$$

For the first term:

$$\frac{\partial}{\partial \tau}\left(\frac{l_i}{\tau}\right) = -\frac{l_i}{\tau^2}.$$

For the second term, using the chain rule:

$$\frac{\partial}{\partial \tau} \log \left(\sum_{j=1}^{K} e^{l_j/\tau} \right) = \frac{1}{\sum_{j=1}^{K} e^{l_j/\tau}} \cdot \frac{\partial}{\partial \tau} \left(\sum_{j=1}^{K} e^{l_j/\tau} \right).$$

The derivative of the exponential sum:

$$\frac{\partial}{\partial \tau} \left(\sum_{j=1}^{K} e^{l_j/\tau} \right) = \sum_{j=1}^{K} e^{l_j/\tau} \cdot (-\frac{l_j}{\tau^2}).$$

Combining these:

$$\frac{\partial}{\partial \tau} \log \left(\sum_{j=1}^{K} e^{l_j/\tau} \right) = -\frac{1}{\tau^2} \cdot \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}.$$

Therefore, the complete first derivative is:

$$\frac{\partial}{\partial \tau} \log p_i(\tau) = \frac{1}{\tau^2} \left(-l_i + \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}} \right).$$

Now for the second derivative.

Defining notation:

$$f(\tau) = \frac{1}{\tau^2}$$

$$g(\tau) = -l_i + \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}$$

Using the product rule:

$$\frac{d}{d\tau}[f(\tau)g(\tau)] = f'(\tau)g(\tau) + f(\tau)g'(\tau).$$

Find $f'(\tau)$:

$$f'(\tau) = \frac{d}{d\tau}(\tau^{-2}) = -2\tau^{-3}$$

For $g'(\tau)$, we need to differentiate the quotient. Let's call:

$$N(\tau) = \sum_{j=1}^{K} l_j e^{l_j/\tau}$$

$$D(\tau) = \sum_{j=1}^{K} e^{l_j/\tau}$$

Then:

$$g'(\tau) = \frac{d}{d\tau}\left(-l_i + \frac{N(\tau)}{D(\tau)}\right).$$

Using the quotient rule:

$$g'(\tau) = \frac{N'(\tau)D(\tau) - N(\tau)D'(\tau)}{[D(\tau)]^2}.$$

Now we need $N'(\tau)$ and $D'(\tau)$:

$$N'(\tau) = \sum_{j=1}^{K} l_j \cdot \frac{d}{d\tau}(e^{l_j/\tau}) = \sum_{j=1}^{K} l_j e^{l_j/\tau}(-l_j/\tau^2)$$

$$D'(\tau) = \sum_{j=1}^{K} \frac{d}{d\tau}(e^{l_j/\tau}) = \sum_{j=1}^{K} e^{l_j/\tau}(-l_j/\tau^2)$$

Putting it all together:

$$\frac{d}{d\tau}\left[\frac{1}{\tau^2}\left(-l_i + \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}\right)\right] =$$

$$-2\tau^{-3}\left(-l_i + \frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}}\right) +$$

$$\tau^{-2}\left(\frac{\sum_{j=1}^{K} -l_j^2\tau^{-2}e^{l_j/\tau}\sum_{k=1}^{K} e^{l_k/\tau} - \sum_{j=1}^{K} l_j e^{l_j/\tau}\sum_{k=1}^{K} -l_k\tau^{-2}e^{l_k/\tau}}{(\sum_{j=1}^{K} e^{l_j/\tau})^2}\right)$$

Simplifying the final expression by factoring out τ^{-2} from the second term's numerator:

$$2\tau^{-3}l_i - 2\tau^{-3}\frac{\sum_{j=1}^{K} l_j e^{l_j/\tau}}{\sum_{j=1}^{K} e^{l_j/\tau}} + \tau^{-4}\left(\frac{-\sum_{j=1}^{K} l_j^2 e^{l_j/\tau}\sum_{k=1}^{K} e^{l_k/\tau} + \sum_{j=1}^{K} l_j e^{l_j/\tau}\sum_{k=1}^{K} l_k e^{l_k/\tau}}{(\sum_{j=1}^{K} e^{l_j/\tau})^2}\right)$$

Let:

$$Z = \sum_{j=1}^{K} e^{l_j/\tau}$$

$$L = \sum_{j=1}^{K} l_j e^{l_j/\tau}$$

$$L_2 = \sum_{j=1}^{K} l_j^2 e^{l_j/\tau}$$

Then our expression becomes:

$$2\tau^{-3}l_i - 2\tau^{-3}\frac{L}{Z} + \tau^{-4}\left(\frac{-L_2 Z + L^2}{Z^2}\right).$$

Combining terms with the same denominator:

$$2\tau^{-3}l_i - 2\tau^{-3}\frac{L}{Z} + \tau^{-4}\left(\frac{L^2 - L_2 Z}{Z^2}\right) = \frac{2\tau^{-3}l_i Z^2 - 2\tau^{-3}LZ + \tau^{-4}(L^2 - L_2 Z)}{Z^2}.$$

Therefore, our final simplified expression is:

$$\frac{2l_i Z^2\tau - 2LZ\tau + L^2 - L_2 Z}{\tau^4 Z^2}, \tag{D.1}$$

where:

$$Z = \sum_{j=1}^{K} e^{l_j/\tau}$$

$$L = \sum_{j=1}^{K} l_j e^{l_j/\tau}$$

$$L_2 = \sum_{j=1}^{K} l_j^2 e^{l_j/\tau}$$

D.3. Full Derivative "Calculations" with SymPy

The calculations above can be quickly recreated using python's SymPy package.

```
from sympy import Symbol, log, exp, Sum, latex, diff

# Define symbols
l_i = Symbol('l_i')  # The i-th logit
tau = Symbol('tau')  # Temperature parameter
j = Symbol('j')      # Summation index
K = Symbol('K')      # Upper bound of summation (vocabulary size)

# Create the expression
expression = l_i/tau - log(Sum(exp(j/tau), (j, 1, K)))

# Calculate first derivative
d1 = diff(expression, tau, 1)

# Calculate second derivative
d2 = diff(expression, tau, 2)
```

Note: l_j is replaced simply by j in the code above. The calculation does not change.

The outputs can be verified with:

```
    print(latex(diff(expression, tau, 1)))
```

Which produces:

$$-\frac{l_i}{\tau^2} - \frac{\sum_{j=1}^{K} -\frac{je^{\frac{j}{\tau}}}{\tau^2}}{\sum_{j=1}^{K} e^{\frac{j}{\tau}}}, \tag{D.2}$$

and:

```
print(latex(diff(expression, tau, 2)))
```

which produces:

$$\frac{2l_i - \frac{\sum_{j=1}^{K} j\left(\frac{j}{\tau}+2\right)e^{\frac{j}{\tau}}}{\sum_{j=1}^{K} e^{\frac{j}{\tau}}} + \frac{\left(\sum_{j=1}^{K} je^{\frac{j}{\tau}}\right)^2}{\tau\left(\sum_{j=1}^{K} e^{\frac{j}{\tau}}\right)^2}}{\tau^3}, \tag{D.3}$$

and which simplifies to equation D.1.

Bibliography

Afshin-Pour, B., M. Qiu, S. Vajargah, et al. (2023). "Discriminating Acute Respiratory Distress Syndrome from other forms of respiratory failure via iterative machine learning". In: *Intelligence-Based Medicine* 7, p. 100087.

Amos, Brandon, Lei Xu, and J. Zico Kolter (2017). "Input Convex Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 146–155.

ARDS Definition Task Force, V Marco Ranieri, Gordon D Rubenfeld, et al. (June 2012). "Acute respiratory distress syndrome: the Berlin Definition". In: *JAMA* 307.23. [Online; accessed 2025-07-07], pp. 2526–33. ISSN: 0098-7484. DOI: 10.1001/jama.2012.5669.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 214–223.

Bastounis, Alexander, Paolo Campodonico, Mihaela van der Schaar, et al. (2024). *On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of 'I don't know'*. arXiv: 2408.02357 [cs.AI]. URL: https://arxiv.org/abs/2408.02357.

Bellani, G., J. Laffey, T. Pham, et al. (2016). "Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries". In: *Jama* 315, pp. 788–800.

Benamou, Jean-David and Yann Brenier (2000). "A Computational Fluid Mechanics Solution to the Monge–Kantorovich Mass Transfer Problem". In: *Numerische Mathematik* 84, pp. 375–393.

Boecking, B., N. Usuyama, S. Bannur, et al. (2022). "Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing". [cs]. arXiv: 2204.09817. URL: http://arxiv.org/abs/2204.09817.

Boffi, Nicholas Matthew, Michael Samuel Albergo, and Eric Vanden-Eijnden (2025). "Flow Map Matching with Stochastic Interpolants: A Mathematical Framework for Consistency Models". In: *Transactions on Machine Learning Research*.

Bourtoule, Lucas, Varun Chandrasekaran, Christopher A Choquette-Choo, et al. (2021). "Machine unlearning". In: *2021 IEEE symposium on security and privacy (SP)*. IEEE, pp. 141–159.

Brenier, Yann (1991). "Polar Factorization and Monotone Rearrangement of Vector-Valued Functions". In: *Communications on Pure and Applied Mathematics* 44.4, pp. 375–417.

Bunne, Charlotte, Stefan G. Stark, Gabriele Gut, et al. (2023). "Learning Single-Cell Perturbation Responses using Neural Optimal Transport". In: *Nature Methods* 20, pp. 1759–1768.

Calfee, Carolyn S, Kevin Delucchi, Polly E Parsons, et al. (Aug. 2014). "Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials". In: *The Lancet. Respiratory medicine* 2.8. [Online; accessed 2025-07-07], pp. 611–20. ISSN: 2213-2600. DOI: 10.1016/S2213-2600(14)70097-9.

Cao, Yinzhi and Junfeng Yang (2015). "Towards making systems forget with machine unlearning". In: *2015 IEEE symposium on security and privacy*. IEEE, pp. 463–480.

Carlini, Nicholas, Florian Tramer, Eric Wallace, et al. (2021). "Extracting training data from large language models". In: *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650.

Chen, T., S. Kornblith, M. Norouzi, et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International Conference On Machine Learning*, pp. 1597–1607.

Courty, Nicolas, Rémi Flamary, Devis Tuia, et al. (2017). "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9, pp. 1853–1865.

Crothers, Evan N., Nathalie Japkowicz, and Herna L. Viktor (2023). "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods". In: *IEEE Access* 11, pp. 70977–71002. DOI: 10.1109/ACCESS.2023.3294090.

Cuturi, Marco (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 26.

Cyphert, Amy B (2023). "Generative AI, plagiarism, and copyright infringement in legal documents". In: *Minn. JL Sci. & Tech.* 25, p. 49.

Deloitte (2026). *State of AI in the Enterprise: 2026 Report*. URL: https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/services/consulting/2026/state-of-ai-2026.pdf.

Dietterich, Thomas G (1998). "Approximate statistical tests for comparing supervised classification learning algorithms". In: *Neural computation* 10.7, pp. 1895–1923.

Dong, Yijiang River, Hongzhou Lin, Mikhail Belkin, et al. (Apr. 2025). "UNDIAL: Self-Distillation with Adjusted Logits for Robust Unlearning in Large Language Models". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 8827–8840. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.444. URL: https://aclanthology.org/2025.naacl-long.444/.

Duarte, André V., Xuandong Zhao, Arlindo L. Oliveira, et al. (2024). "DE-COP: detecting copyrighted content in language models training data". In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org.

Eldan, Ronen and Mark Russinovich (2024). *Who's Harry Potter? Approximate Unlearning for LLMs*. URL: https://openreview.net/forum?id=PDct7vrcvT.

Entesari, Taha, Arman Hatami, Rinat Khaziev, et al. (2026). "Constrained Entropic Unlearning: A Primal-Dual Framework for Large Language Models". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=ZtB34bQI54.

Erol, Tolga, Arif Furkan Mendi, and Dilara Doğan (2020). "The digital twin revolution in healthcare". In: *2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*. IEEE, pp. 1–7.

Fatras, Kilian, Younes Zine, Rémi Flamary, et al. (2019). "Learning with Minibatch Wasserstein: Asymptotic and Gradient Properties". In: *arXiv preprint arXiv:1910.04091*.

Ferguson, Niall D, Robert M Kacmarek, Jean-Daniel Chiche, et al. (June 2004). "Screening of ARDS patients using standardized ventilator settings: influence on enrollment in a clinical trial". In:

Intensive care medicine 30.6. [Online; accessed 2025-07-07], pp. 1111–6. ISSN: 0342-4642. DOI: 10.1007/s00134-004-2163-2.

Fletcher, R. (2000). *Practical Methods of Optimization*. Wiley. ISBN: 9780471494638. URL: https://books.google.com/books?id=z3m_EAAAQBAJ.

Frans, Kevin, Danijar Hafner, Sergey Levine, et al. (2024). "One Step Diffusion via Shortcut Models". In: *arXiv preprint arXiv:2410.12557*.

Freeman, Joshua, Chloe Rippe, Edoardo Debenedetti, et al. (2024). "Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 Lawsuit". In: *Neurips Safe Generative AI Workshop 2024*. URL: https://openreview.net/forum?id=C66DBl9At8.

Fu, Wenjie, Huandong Wang, Chen Gao, et al. (2023). "Practical membership inference attacks against fine-tuned large language models via self-prompt calibration". In: *arXiv preprint arXiv:2311.06062*.

Gandomi, A., P. Wu, D. Clement, et al. (2024). "ARDSFlag: an NLP/machine learning algorithm to visualize and detect high-probability ARDS admissions independent of provider recognition and billing codes". In: *BMC Medical Informatics And Decision Making* 24, p. 195.

Gao, Chongyang, Lixu Wang, Chenkai Weng, et al. (2024). "Practical unlearning for large language models". In: *arXiv e-prints*, arXiv–2407.

Gao, Leo, Jonathan Tow, Baber Abbasi, et al. (July 2024). *The Language Model Evaluation Harness*. Version v0.4.3. DOI: 10.5281/zenodo.12608602. URL: https://zenodo.org/records/12608602.

Gao, Zijun, Shuyang Cheng, Emily Wittrup, et al. (2024). "Learning using privileged information with logistic regression on acute respiratory distress syndrome detection". In: *Artificial Intelligence in Medicine* 156, p. 102947. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2024.102947. URL: https://www.sciencedirect.com/science/article/pii/S0933365724001891.

Geng, Zhengyang, Mingyang Deng, Xingjian Bai, et al. (2025). "Mean Flows for One-Step Generative Modeling". In: *arXiv preprint arXiv:2505.13447*.

Ginart, Antonio, Melody Guan, Gregory Valiant, et al. (2019). "Making AI forget you: Data deletion in machine learning". In: *Advances in neural information processing systems* 32.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Grynbaum, Michael M and Ryan Mac (2023). "The Times sues OpenAI and Microsoft over AI use of copyrighted work". In: *The New York Times* 27.1.

Gundersen, Gregory (Nov. 2019). *Asymptotic Normality of Maximum Likelihood Estimators*. URL: https://gregorygundersen.com/blog/2019/11/28/asymptotic-normality-mle/.

Guo, Chuan, Tom Goldstein, Awni Hannun, et al. (2019). "Certified data removal from machine learning models". In: *arXiv preprint arXiv:1911.03030*.

Gushchin, Nikita, Sergei Kholkin, Evgeny Burnaev, et al. (2024). "Light and Optimal Schrödinger Bridge Matching". In: *Proceedings of the 41st International Conference on Machine Learning*. arXiv:2402.03207. URL: https://arxiv.org/abs/2402.03207.

Gustafsson, Stefan, Daniel Gedon, Erik Lampa, et al. (Nov. 2022). "Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients". In: *Scientific Reports* 12.1. ISSN: 2045-2322. DOI: 10.1038/s41598-022-24254-x. URL: http://dx.doi.org/10.1038/s41598-022-24254-x.

Hannun, Awni Y., Pranav Rajpurkar, Masoumeh Haghpanahi, et al. (Jan. 2019). "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network". In: *Nature Medicine* 25.1, pp. 65–69. ISSN: 1078-8956. DOI: 10.1038/s41591-018-0268-3. URL: http://dx.doi.org/10.1038/s41591-018-0268-3.

Hassija, Vikas, Vinay Chamola, Atmesh Mahapatra, et al. (2024). "Interpreting black-box models: a review on explainable artificial intelligence". In: *Cognitive Computation* 16.1, pp. 45–74.

Hatami, Mohsen, Van Tuan Pham, Hozefa Lakadawala, et al. (2026). *Securing AI Agents in Cyber-Physical Systems: A Survey of Environmental Interactions, Deepfake Threats, and Defenses*. arXiv: 2601.20184 [cs.CR]. URL: https://arxiv.org/abs/2601.20184.

He, K., X. Zhang, S. Ren, et al. (2016). "Deep residual learning for image recognition". In: *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*, pp. 770–778.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015). *Distilling the Knowledge in a Neural Network*. arXiv: 1503.02531 [stat.ML]. URL: https://arxiv.org/abs/1503.02531.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics*, pp. 65–70.

Holtzman, Ari, Jan Buys, Li Du, et al. (2020). *The Curious Case of Neural Text Degeneration*. arXiv: 1904.09751 [cs.CL]. URL: https://arxiv.org/abs/1904.09751.

Holtzman, Ari, Jan Buys, Maxwell Forbes, et al. (2018). *Learning to Write with Cooperative Discriminators*. arXiv: 1805.06087 [cs.CL]. URL: https://arxiv.org/abs/1805.06087.

Hong, Yihuai, Yuelin Zou, Lijie Hu, et al. (2024). "Dissecting fine-tuning unlearning in large language models". In: *arXiv preprint arXiv:2410.06606*.

Hu, Ruihan, Yu-Ming Shang, Jiankun Peng, et al. (2025). "Automated Detection of Pre-training Text in Black-box LLMs". In: *arXiv preprint arXiv:2506.19399*.

Huang, Yuheng, Jiayang Song, Zhijie Wang, et al. (Feb. 2025). "¡i¿Look Before You Leap:¡/i¿ An Exploratory Study of Uncertainty Analysis for Large Language Models". In: *IEEE Transactions on Software Engineering* 51.2, pp. 413–429. ISSN: 2326-3881. DOI: 10.1109/tse.2024.3519464. URL: http://dx.doi.org/10.1109/TSE.2024.3519464.

Huh, Minyoung, Brian Cheung, Tongzhou Wang, et al. (2024). *The Platonic Representation Hypothesis*. arXiv: 2405.07987 [cs.LG]. URL: https://arxiv.org/abs/2405.07987.

Hutson, Jevan, Cedric Whitney, and Jay T. Conrad (Jan. 2026). "Forget Me Not? Machine Unlearning's Implication for Privacy Law". In: *Science and Technology Law Review* 27.1. DOI: 10.52214/stlr.v27i1.14547. URL: https://journals.library.columbia.edu/index.php/stlr/article/view/14547.

Ilharco, Gabriel, Marco Tulio Ribeiro, Mitchell Wortsman, et al. (2022). "Editing models with task arithmetic". In: *arXiv preprint arXiv:2212.04089*.

Ippolito, Daphne, Nicholas Carlini, Katherine Lee, et al. (2023). *Reverse-Engineering Decoding Strategies Given Blackbox Access to a Language Generation System*. arXiv: 2309.04858 [cs.LG]. URL: https://arxiv.org/abs/2309.04858.

Ishibashi, Yoichi and Hidetoshi Shimodaira (2024). *Knowledge Sanitization of Large Language Models*. arXiv: 2309.11852 [cs.CL]. URL: https://arxiv.org/abs/2309.11852.

J.K. Rowling (2013). *Harry Potter and the Prisoner of Azkaban*. Harry Potter. Scholastic. ISBN: 9780545582933. URL: https://books.google.com/books?id=uUp5mAEACAAJ.

Jabbour, Sarah, David Fouhey, Ella Kazerooni, et al. (May 2022). "Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure". In: *Journal of the American Medical Informatics Association : JAMIA* 29.6. [Online; accessed 2025-07-07], pp. 1060–1068. ISSN: 1067-5027. DOI: 10.1093/jamia/ocac030.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Jim (Aug. 2025). *Temperature in GPT-5 models*. URL: https://community.openai.com/t/temperature-in-gpt-5-models/1337133.

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, et al. (2025). *Why Language Models Hallucinate*. arXiv: 2509.04664 [cs.CL]. URL: https://arxiv.org/abs/2509.04664.

Karamolegkou, Antonia, Jiaang Li, Li Zhou, et al. (2023). "Copyright violations and large language models". In: *arXiv preprint arXiv:2310.13771*.

Kärkkäinen, Kimmo and Jungseock Joo (2019). "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age". In: *CoRR* abs/1908.04913. arXiv: 1908.04913. URL: http://arxiv.org/abs/1908.04913.

Kempton, Tom, Stuart Burrell, and Connor Cheverall (2025). *TempTest: Local Normalization Distortion and the Detection of Machine-generated Text*. arXiv: 2503.20421 [cs.CL]. URL: https://arxiv.org/abs/2503.20421.

Kim, Dongjun, Chieh-Hsin Lai, Wei-Hsiang Liao, et al. (2023). "Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion". In: *arXiv preprint arXiv:2310.02279*.

Kim, Hyo Seo, Dongyoon Han, and Junsuk Choe (2025). "NegMerge: Sign-Consensual Weight Merging for Machine Unlearning". In: *Forty-second International Conference on Machine Learning*. URL: https://openreview.net/forum?id=ZbWXovStjD.

Kong, Lingkai, Molei Tao, Yang Liu, et al. (2026). "AlignFlow: Improving Flow-Based Generative Models with Semi-Discrete Optimal Transport". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=nTCF3QNsIN.

Kornilov, Nikita, Petr Mokrov, Alexander Gasnikov, et al. (2024). "Optimal Flow Matching: Learning Straight Trajectories in Just One Step". In: *Advances in Neural Information Processing Systems*. arXiv:2403.13117. URL: https://arxiv.org/abs/2403.13117.

Korotin, Alexander, Daniil Selikhanovych, and Evgeny Burnaev (2023). "Neural Optimal Transport". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=d8CBRlWNkqH.

LangChain (2026). URL: https://www.langchain.com/resources/llm-observability-tools.

Lehmann, E.L. and G. Casella (2006). *Theory of Point Estimation*. Springer Texts in Statistics. Springer New York. ISBN: 9780387227283. URL: https://books.google.com/books?id=4f24CgAAQBAJ.

Levy, E., D. Claar, B. Fuchs, et al. (2025). "Development and External Validation of a Detection Model to Retrospectively Identify Patients With Acute Respiratory Distress Syndrome". In: *Critical Care Medicine*, pp. 10–1097.

Li, Xiaomin, Zhou Yu, Ziji Zhang, et al. (2025). *Semantic Volume: Quantifying and Detecting both External and Internal Uncertainty in LLMs*. arXiv: 2502.21239 [cs.CL]. URL: https://arxiv.org/abs/2502.21239.

Lin, Zhen, Shubhendu Trivedi, and Jimeng Sun (2024). *Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models*. arXiv: 2305.19187 [cs.CL]. URL: https://arxiv.org/abs/2305.19187.

Lipman, Yaron, Ricky T. Q. Chen, Heli Ben-Hamu, et al. (2022). "Flow Matching for Generative Modeling". In: *arXiv preprint arXiv:2210.02747*.

Liu, Bo, Qiang Liu, and Peter Stone (2022). "Continual learning and private unlearning". In: *Conference on Lifelong Learning Agents*. PMLR, pp. 243–254.

Liu, Qiang et al. (2024). *Rectified Flow: Tutorial and Resources*. https://rectifiedflow.github.io/.

Liu, Sijia, Yuanshun Yao, Jinghan Jia, et al. (2025). "Rethinking machine unlearning for large language models". In: *Nature Machine Intelligence*, pp. 1–14.

Liu, Xiaoou, Tiejin Chen, Longchao Da, et al. (Aug. 2025). "Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey". In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. KDD '25. ACM, pp. 6107–6117. DOI: 10.1145/3711896.3736569. URL: http://dx.doi.org/10.1145/3711896.3736569.

Liu, Xingchao, Chengyue Gong, and Qiang Liu (2023). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *International Conference on Learning Representations*. arXiv:2209.03003. URL: https://arxiv.org/abs/2209.03003.

Liu, Ye, Rui Meng, Shafiq Joty, et al. (2025). *CodeXEmbed: A Generalist Embedding Model Family for Multiligual and Multi-task Code Retrieval*. arXiv: 2411.12644 [cs.SE]. URL: https://arxiv.org/abs/2411.12644.

Liu, Zheyuan, Guangyao Dou, Zhaoxuan Tan, et al. (2024). "Towards safer large language models through machine unlearning". In: *arXiv preprint arXiv:2402.10058*.

Liu, Zuxin, Thai Hoang, Jianguo Zhang, et al. (2024). *APIGen: Automated Pipeline for Generating Verifiable and Diverse Function-Calling Datasets*. arXiv: 2406.18518 [cs.CL]. URL: https://arxiv.org/abs/2406.18518.

Lopez, V., A. Fernandez, S. Garcia, et al. (2013). "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". In: *Information Sciences* 250, pp. 113–141.

Ma, Chenrui, Xi Xiao, Tianyang Wang, et al. (2026). *Stochastic Interpolants via Conditional Dependent Coupling*. OpenReview preprint. URL: https://openreview.net/forum?id=7FZFkJKD9f.

Maini, Pratyush, Zhili Feng, Avi Schwarzschild, et al. (2024). "Tofu: A task of fictitious unlearning for llms". In: *arXiv preprint arXiv:2401.06121*.

Makkuva, Ashok Vardhan, Amirhossein Taghvaei, Sewoong Oh, et al. (2020). "Optimal Transport Mapping via Input Convex Neural Networks". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 6672–6681.

Marks, Samuel and Max Tegmark (2024). *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*. arXiv: 2310.06824 [cs.AI]. URL: https://arxiv.org/abs/2310.06824.

Matthay, M., R. Zemans, G. Zimmerman, et al. (2019). "Acute respiratory distress syndrome". In: *Nature Reviews Disease Primers* 5, p. 18.

Matthay, Michael A., Yaseen Arabi, Alejandro C. Arroliga, et al. (Jan. 2024). "A New Global Definition of Acute Respiratory Distress Syndrome". In: *American Journal of Respiratory and*

Critical Care Medicine 209.1, pp. 37–47. ISSN: 1073-449X. DOI: 10.1164/rccm.202303-0558WS. URL: https://doi.org/10.1164/rccm.202303-0558WS.

McCann, Robert J (1997). "A convexity principle for interacting gases". In: *Advances in mathematics* 128.1, pp. 153–179.

Meng, Kevin, David Bau, Alex Andonian, et al. (2022). "Locating and editing factual associations in gpt". In: *Advances in neural information processing systems* 35, pp. 17359–17372.

Meng, Kevin, Arnab Sen Sharma, Alex Andonian, et al. (2023). "Mass-Editing Memory in a Transformer". In: *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

Metz, Cade (2024). "OpenAI Says New York Times Lawsuit Against It Is 'Without Merit'." In: *The New York Times (Digital Edition)*, NA–NA.

Microsoft, : Abdelrahman Abouelenin, et al. (2025). *Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs*. arXiv: 2503.01743 [cs.CL]. URL: https://arxiv.org/abs/2503.01743.

Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, et al. (July 2023). "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 24950–24962. URL: https://proceedings.mlr.press/v202/mitchell23a.html.

Mousavi-Hosseini, Alireza, Stephen Y. Zhang, Michal Klein, et al. (2025). "Flow Matching with Semidiscrete Couplings". In: *arXiv preprint arXiv:2509.25519*. URL: https://arxiv.org/abs/2509.25519.

Nadeau, Claude and Yoshua Bengio (2003). "Inference for the generalization error". In: *Machine learning* 52.3, pp. 239–281.

Nagrani, Arsha, Shan Yang, Anurag Arnab, et al. (2021). "Attention bottlenecks for multimodal fusion". In: *Advances in neural information processing systems* 34, pp. 14200–14213.

Naseh, Ali, Kalpesh Krishna, Mohit Iyyer, et al. (2023). "Stealing the Decoding Algorithms of Language Models". In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and*

Communications Security. CCS '23. Copenhagen, Denmark: Association for Computing Machinery, pp. 1835–1849. ISBN: 9798400700507. DOI: 10.1145/3576915.3616652. URL: https://doi.org/10.1145/3576915.3616652.

Needham, Dale M, Ting Yang, Victor D Dinglas, et al. (Jan. 2015). "Timing of low tidal volume ventilation and intensive care unit mortality in acute respiratory distress syndrome. A prospective cohort study". In: *American journal of respiratory and critical care medicine* 191.2. [Online; accessed 2025-07-07], pp. 177–85. ISSN: 1073-449X. DOI: 10.1164/rccm.201409-1598OC.

Orgad, Hadas, Michael Toker, Zorik Gekhman, et al. (2025). *LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations*. arXiv: 2410.02707 [cs.CL]. URL: https://arxiv.org/abs/2410.02707.

Pai, K., W. Chao, Y. Huang, et al. (2022). "Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs". In: *Digital Health* 8. p 76221120317 (2022,8), p. 205520. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9386858/.

Pan, Melissa Z., Negar Arabzadeh, Riccardo Cogo, et al. (2026). *Measuring Agents in Production*. arXiv: 2512.04123 [cs.CY]. URL: https://arxiv.org/abs/2512.04123.

Patil, Shishir G., Tianjun Zhang, Xin Wang, et al. (2023). *Gorilla: Large Language Model Connected with Massive APIs*. arXiv: 2305.15334 [cs.CL]. URL: https://arxiv.org/abs/2305.15334.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peyré, Gabriel and Marco Cuturi (2019). "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5–6, pp. 355–607.

Pooladian, Aram-Alexandre, Heli Ben-Hamu, Carles Domingo-Enrich, et al. (2023). "Multisample Flow Matching: Straightening Flows with Minibatch Couplings". In: *arXiv preprint arXiv:2304.14772*.

Puerto, Haritz, Martin Gubri, Sangdoo Yun, et al. (2025). "Scaling up membership inference: When and how attacks succeed on large language models". In: *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182.

Rabanser, Stephan, Sayash Kapoor, Peter Kirgis, et al. (2026). *Towards a Science of AI Agent Reliability*. arXiv: 2602.16666 [cs.AI]. URL: https://arxiv.org/abs/2602.16666.

Rabinovich, Ella and Ateret Anaby Tavor (May 2025). "On the Robustness of Agentic Function Calling". In: *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*. Ed. by Trista Cao, Anubrata Das, Tharindu Kumarage, et al. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 298–304. ISBN: 979-8-89176-233-6. DOI: 10.18653/v1/2025.trustnlp-main.20. URL: https://aclanthology.org/2025.trustnlp-main.20/.

Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, et al. (2023). "Direct preference optimization: Your language model is secretly a reward model". In: *Advances in neural information processing systems* 36, pp. 53728–53741.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, et al. (2017). "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *CoRR* abs/1711.05225. arXiv: 1711.05225. URL: http://arxiv.org/abs/1711.05225.

Ranieri, V., G. Rubenfeld, B. Taylor Thompson, et al. (2012). "Acute respiratory distress syndrome: the Berlin Definition". In: *JAMA: Journal Of The American Medical Association* 307.

Ray, Partha Pratim (2026). "A comprehensive introspection on AI risks: taxonomy, challenges, and future directions". In: *Iran Journal of Computer Science* 9.1, p. 18.

Reamaroon, N., M. Sjoding, J. Gryak, et al. (2021). "Automated detection of acute respiratory distress syndrome from chest X-Rays using Directionality Measure and deep learning features". In: *Computers In Biology And Medicine* 134, p. 104463.

Reamaroon, Narathip, Michael W Sjoding, Kaiwen Lin, et al. (Jan. 2019). "Accounting for Label Uncertainty in Machine Learning for Detection of Acute Respiratory Distress Syndrome". In: *IEEE journal of biomedical and health informatics* 23.1. [Online; accessed 2025-07-07], pp. 407–415. ISSN: 2168-2194. DOI: 10.1109/JBHI.2018.2810820.

Rehm, G., I. Cortés-Puch, B. Kuhn, et al. (2021). "Use of machine learning to screen for acute respiratory distress syndrome using raw ventilator waveform data". In: *Critical Care Explorations* 3, e0313.

Rowling, Joanne K (2023). *Harry Potter and the sorcerer's stone*. Scholastic Incorporated.

Rubulotta, Francesca, Sahar Bahrami, Dominic C Marshall, et al. (2024). "Machine Learning Tools for Acute Respiratory Distress Syndrome Detection and Prediction". In: *Critical Care Medicine* 52.11, pp. 1768–1780.

Rudin, Cynthia (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. arXiv: 1811.10154 [stat.ML]. URL: https://arxiv.org/abs/1811.10154.

Saito, Takaya and Marc Rehmsmeier (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3, e0118432.

Salimans, Tim and Jonathan Ho (2022). "Progressive Distillation for Fast Sampling of Diffusion Models". In: *arXiv preprint arXiv:2202.00512*.

Santambrogio, Filippo (2015). *Optimal Transport for Applied Mathematicians*. Birkhäuser.

Sarah, Wynn-Williams (2025). *Careless people: A cautionary tale of power, greed, and lost idealism*.

Schmid, F., K. Koutini, and G. Widmer (2023a). "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation". In: *ICASSP 2023-2023 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pp. 1–5.

— (2023b). "Low-complexity audio embedding extractors". In: *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 451–455.

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117.

Schneider, Sinclair, Florian Steuber, Joao A. G. Schneider, et al. (2023). *How well can machine-generated texts be identified and can language models be trained to avoid identification?* arXiv: 2310.16992 [cs.CL]. URL: https://arxiv.org/abs/2310.16992.

Seitzer, Maximilian (Aug. 2020). *pytorch-fid: FID Score for PyTorch*. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Sharma, Raghav and Manan Mehta (2025). *Small Language Models for Agentic Systems: A Survey of Architectures, Capabilities, and Deployment Trade offs*. arXiv: 2510.03847 [cs.AI]. URL: https://arxiv.org/abs/2510.03847.

Shi, Shaojie, Xiaoyu Tan, Xihe Qiu, et al. (2024). "Ulmr: Unlearning large language models via negative response and model parameter average". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 755–762.

Shi, Weijia, Anirudh Ajith, Mengzhou Xia, et al. (2024). *Detecting Pretraining Data from Large Language Models*. arXiv: 2310.16789 [cs.CL]. URL: https://arxiv.org/abs/2310.16789.

Shorinwa, Ola, Zhiting Mei, Justin Lidard, et al. (Sept. 2025). "A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions". In: *ACM Computing Surveys* 58.3, pp. 1–38. ISSN: 1557-7341. DOI: 10.1145/3744238. URL: http://dx.doi.org/10.1145/3744238.

Sjoding, M., D. Taylor, J. Motyka, et al. (2021). "Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation". In: *The Lancet Digital Health* 3. Publisher: Elsevier. (2021,6), e340–e348. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00056-X/fulltext.

Sjoding, Michael W, Timothy P Hofer, Ivan Co, et al. (Feb. 2019). "Differences between Patients in Whom Physicians Agree and Disagree about the Diagnosis of Acute Respiratory Distress Syndrome". In: *Annals of the American Thoracic Society* 16.2. [Online; accessed 2025-07-07], pp. 258–264. ISSN: 2325-6621. DOI: 10.1513/AnnalsATS.201806-434OC.

Song, Yang, Prafulla Dhariwal, Mark Chen, et al. (2023). "Consistency Models". In: *arXiv preprint arXiv:2303.01469*.

Sun, Tianze, Xiwang He, and Zhonghai Li (2023). "Digital twin in healthcare: Recent updates and challenges". In: *Digital health* 9, p. 20552076221149651.

Tay, Yi, Dara Bahri, Che Zheng, et al. (2020). *Reverse Engineering Configurations of Neural Text Generation Models*. arXiv: 2004.06201 [cs.CL]. URL: https://arxiv.org/abs/2004.06201.

Team, Gemma, Thomas Mesnard, Cassidy Hardin, et al. (2024). *Gemma: Open Models Based on Gemini Research and Technology*. arXiv: 2403.08295 [cs.CL]. URL: https://arxiv.org/abs/2403.08295.

Thille, Arnaud W, Andrés Esteban, Pilar Fernández-Segoviano, et al. (Apr. 2013). "Comparison of the Berlin definition for acute respiratory distress syndrome with autopsy". In: *American*

journal of respiratory and critical care medicine 187.7. [Online; accessed 2025-07-07], pp. 761–7. ISSN: 1073-449X. DOI: 10.1164/rccm.201211-1981OC.

Tjepkema-Cloostermans, Marleen C, Catarina da Silva Lourenço, Barry J Ruijter, et al. (Oct. 2019). "Outcome Prediction in Postanoxic Coma With Deep Learning". In: *Critical care medicine* 47.10. [Online; accessed 2025-07-12], pp. 1424–1432. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000003854.

Tong, Alexander, Kilian Fatras, Nikolay Malkin, et al. (2023). "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport". In: *arXiv preprint arXiv:2302.00482*.

Valda Toro, Patricia L, Andrew Willmore, Nelson E Wu, et al. (Apr. 2024). "Rapidly improving ARDS differs clinically and biologically from persistent ARDS". In: *Critical care (London, England)* 28.1. [Online; accessed 2025-07-07], p. 132. ISSN: 1364-8535. DOI: 10.1186/s13054-024-04883-6.

Villani, Cédric (2009). *Optimal Transport: Old and New*. Springer.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, Lei, Chen Ma, Xueyang Feng, et al. (Mar. 2024). "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: http://dx.doi.org/10.1007/s11704-024-40231-1.

Wang, Weixuan, Barry Haddow, Alexandra Birch, et al. (2023). *Assessing the Reliability of Large Language Model Knowledge*. arXiv: 2310.09820 [cs.CL]. URL: https://arxiv.org/abs/2310.09820.

Wang, Xiaosong, Yifan Peng, Le Lu, et al. (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Wang, Zhijie, Yuheng Huang, Lei Ma, et al. (2022). *An Exploratory Study of AI System Risk Assessment from the Lens of Data Distribution and Uncertainty*. arXiv: 2212.06828 [cs.LG]. URL: https://arxiv.org/abs/2212.06828.

Wataoka, Koki, Tsubasa Takahashi, and Ryokan Ri (2025). *Self-Preference Bias in LLM-as-a-Judge*. arXiv: 2410.21819 [cs.CL]. URL: https://arxiv.org/abs/2410.21819.

Weiss, C., D. Baker, S. Weiner, et al. (2016). "Low tidal volume ventilation use in acute respiratory distress syndrome". In: *Critical Care Medicine* 44, pp. 1515–1522.

Weng, Xinyu, Nan Zhuang, Jingjing Tian, et al. (2017). *CheXNet for Classification and Localization of Thoracic Diseases*. GitHub. URL: https://github.com/arnoweng/CheXNet.

Wolf, Thomas, Lysandre Debut, Victor Sanh, et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL]. URL: https://arxiv.org/abs/1910.03771.

Xiong, Yaxin, Yuan Gao, Yucheng Qi, et al. (Jan. 2025). "Accuracy of artificial intelligence algorithms in predicting acute respiratory distress syndrome: a systematic review and meta-analysis". In: *BMC medical informatics and decision making* 25.1. [Online; accessed 2025-07-07], p. 44. ISSN: 1472-6947. DOI: 10.1186/s12911-025-02869-0.

Xu, Haoming, Shuxun Wang, Yanqiu Zhao, et al. (July 2025). "ZJUKLAB at SemEval-2025 Task 4: Unlearning via Model Merging." In: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Ed. by Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, et al. Vienna, Austria: Association for Computational Linguistics, pp. 566–574. ISBN: 979-8-89176-273-2. URL: https://aclanthology.org/2025.semeval-1.79/.

Xu, Shizhou and Thomas Strohmer (2023). "Fair data representation for machine learning at the Pareto frontier". In: *Journal of Machine Learning Research* 24.331, pp. 1–63.

— (2025). "Machine Unlearning via Information Theoretic Regularization". In: *arXiv preprint arXiv:2502.05684*.

Yang, An, Baosong Yang, Beichen Zhang, et al. (2024). "Qwen2.5 Technical Report". In: *arXiv preprint arXiv:2412.15115*.

Yang, Karren Dai, Anastasiya Belyaeva, Saradha Venkatachalapathy, et al. (2021). "Multi-domain translation between single-cell imaging and sequencing data using autoencoders". In: *Nature communications* 12.1, p. 31.

Yao, Shunyu, Noah Shinn, Pedram Razavi, et al. (2024). *τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains*. arXiv: 2406.12045 [cs.AI]. URL: https://arxiv.org/abs/2406.12045.

Yao, Yuanshun, Xiaojun Xu, and Yang Liu (2024). "Large language model unlearning". In: *Advances in Neural Information Processing Systems* 37, pp. 105425–105475.

Ye, Jiayi, Yanbo Wang, Yue Huang, et al. (2025). "Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge". In: *International Conference on Learning Representations*. Ed. by Y. Yue, A. Garg, N. Peng, et al. Vol. 2025, pp. 102351–102390. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/fdca08d371e4b6c031397909e20043bd-Paper-Conference.pdf.

Yoon, J., Y. Zhang, J. Jordon, et al. (2020). "Vime: Extending the success of self-and semi-supervised learning to tabular domain". In: *Advances In Neural Information Processing Systems* 33, pp. 11033–11043.

Zeiberg, Daniel, Tejas Prahlad, Brahmajee K Nallamothu, et al. (Mar. 2019). "Machine learning for patient risk stratification for acute respiratory distress syndrome". In: *PloS one* 14.3. [Online; accessed 2025-07-07], e0214465. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0214465.

Zhang, Ruiqi, Licong Lin, Yu Bai, et al. (2024a). *Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning*. arXiv: 2404.05868 [cs.LG]. URL: https://arxiv.org/abs/2404.05868.

— (2024b). "Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning". In: *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=MXLBXjQkmb.

Zhang, Weichao, Ruqing Zhang, Jiafeng Guo, et al. (2024). "Pretraining data detection for large language models: A divergence-based calibration method". In: *arXiv preprint arXiv:2409.14781*.

Zhang, Y., H. Jiang, Y. Miura, et al. (2022). "Contrastive Learning of Medical Visual Representations from Paired Images and Text". [cs]. arXiv, 9). arXiv: 2010.00747. URL: http://arxiv.org/abs/2010.00747.

Zhang, Yichi, Yici Yan, Alex Schwing, et al. (2025). "Hierarchical Rectified Flow Matching with Mini-Batch Couplings". In: *arXiv preprint arXiv:2507.13350*. URL: https://arxiv.org/abs/2507.13350.

Zheng, Wei-Long, Edilberto Amorim, Jin Jing, et al. (Dec. 2021). "Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks". In: *Resuscitation* 169. [Online; accessed 2025-07-11], pp. 86–94. ISSN: 0300-9572. DOI: 10.1016/j.resuscitation.2021.10.034.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, et al. (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.

Zipf, G.K. (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin. URL: https://books.google.com/books?id=y4waAAAAIAAJ.

Zou, Haolin, Arnab Auddy, Yongchan Kwon, et al. (2025). *Certified Data Removal Under High-dimensional Settings*. arXiv: 2505.07640 [stat.ML]. URL: https://arxiv.org/abs/2505.07640.