# CONTROLLING AI TRUSTWORTHINESS THROUGH DATA AND PROMPT SELECTION

*Shubhadip Nag*

# CONTROLLING AI TRUSTWORTHINESS THROUGH DATA AND PROMPT SELECTION

*Thesis submitted to the*
*Indian Institute of Technology, Kharagpur*
*For award of the degree*

*of*

**Master of Science**

*by*

**Shubhadip Nag**

**Under the supervision of**

**Prof. Sourangshu Bhattacharya**


**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR**

**January 2026**

# APPROVAL OF THE VIVA-VOCE BOARD

Date: / / 20

Certified that the thesis entitled **Controlling AI Trustworthiness through Data and Prompt selection** submitted by Shubhadip Nag to the Indian Institute of Technology, Kharagpur, for the award of the degree of Master of Science has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

(Member of DAC) (Member of DAC) (Member of DAC)

(Supervisor)

(External Examiner) (Chairman)

# CERTIFICATE

*This is to certify that the thesis entitled* ***Controlling AI Trustworthiness through Data and Prompt selection****, submitted by Shubhadip Nag to the Indian Institute of Technology, Kharagpur, for the partial fulfillment of the award of the degree of Master of Science in Computer Science and Engineering, is a record of bona fide research work carried out by him under our supervision and guidance.*

*The thesis in our opinion, is worthy of consideration for the award of the degree of Master of Science in accordance with the regulations of the institute. To the best of our knowledge, the results embodied in this thesis have not been submitted to any other university or institute for the award of any other degree or diploma.*

Sourangshu Bhattacharya
Associate Professor
Computer Science and Engineering,
IIT Kharagpur

Date:

# DECLARATION

I certify that

a. the work contained in this thesis is original and has been done by me under the guidance of my supervisors.

b. the work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in preparing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Shubhadip Nag

# ACKNOWLEDGMENTS

I am profoundly indebted to my supervisor, Prof. Sourangshu Bhattacharya, for his unwavering guidance, constant mentorship, and emphasis on independent, critical, and rigorous thinking. His deep technical insights, constructive feedback, and high research standards consistently challenged me to refine my ideas and approach problems with clarity and precision. Beyond academic guidance, his encouragement and trust played a pivotal role in shaping my research mindset and overall growth as a researcher throughout my MS.

I express my deepest gratitude to Maa and Baba for their unwavering support, encouragement, and belief in my decision to pursue a research degree. I am equally grateful to my Mama and Mimi, and to my sisters, whose constant care, motivation, and sacrifices provided me with strength and reassurance throughout this journey and made this achievement possible.

I sincerely thank Chinu Sir, my first teacher of Computer Science, who introduced me to the fundamentals of the subject and ignited my early interest in computer science and research.

I am grateful to Prof. Niloy Ganguly for giving me the opportunity to be part of the Complex Networks Research Group (CNeRG). Through CNeRG, I was exposed to high-quality research, leading AI conferences, and a vibrant academic environment that significantly strengthened my understanding of research methodologies and scientific communication.

I am especially thankful to Dr. Soumi Das for her mentorship, encouragement,

and guidance during my MS. She introduced me to the research direction of data-centric AI and motivated me to continue exploring this area, which became a core component of my thesis.

I would also like to acknowledge my collaborators Dr. Suparna Bhattacharya and Tarun Kumar from HP Labs for their valuable discussions, insights, and collaborative support, which greatly enriched the research presented in this thesis.

I sincerely thank all my co-authors for their collaborative spirit, technical discussions, and constructive feedback, which were instrumental in improving the quality of the published work arising from this thesis.

I also acknowledge the CNeRG weekly reading group for broadening my understanding of the research landscape and sharpening my ability to critically analyze research papers. Finally, I thank the members of CNeRG and the Department of Computer Science and Engineering at IIT Kharagpur for providing a supportive and intellectually stimulating research environment.

Shubhadip Nag

IIT Kharagpur, India

# ABSTRACT

The deployment of AI systems in real-world, high-stakes applications demands strong guarantees on trustworthiness, including fairness, robustness, and safety. However, modern learning systems often exhibit undesirable behaviors such as bias, brittleness, or leakage of sensitive concepts, arising from a small but influential subset of training signals. This thesis investigates *selection* as a principled and general mechanism for controlling AI trustworthiness by carefully choosing what the model learns from.

The first part of this thesis focuses on classical supervised learning and proposes **VTruST**, a data-centric framework for constructing trustworthy training datasets. VTruST formulates data subset selection as a controllable value-function optimization problem, enabling explicit trade-offs between accuracy, fairness, and robustness. By posing data valuation as an online sparse approximation problem and introducing an efficient online Orthogonal Matching Pursuit–based selection algorithm, VTruST identifies high-impact training samples that govern trustworthiness outcomes. Extensive experiments across social, image, and scientific datasets demonstrate that models trained on VTruST-selected subsets outperform state-of-the-art baselines while also providing data-centric explanations for trust-related behaviors.

The second part of the thesis extends the selection-centric perspective to large language models (LLMs), where trust violations often manifest through prompt-dependent exposure of biased or harmful concepts. We introduce **MPSelectTune**, a prompt-type selection framework for robust concept unlearning in LLMs. MPSelectTune adopts a two-stage fine-tuning strategy that first leverages multiple prompt

types and then identifies and fine-tunes on the worst-case prompt type, i.e., the prompt that most effectively elicits the sensitive concept. This adversarial prompt selection significantly reduces worst-case concept leakage while preserving main task performance across diverse prompt variations. Experimental results on multiple benchmarks and LLM architectures show consistent improvements over existing unlearning methods in both average and worst-case settings.

Together, these contributions establish the selection of training signals, data points in supervised learning, and prompt types in LLMs as a unifying and effective paradigm for controlling trustworthiness in AI systems.

**Keywords:** Trustworthy AI, Data-Centric AI, Dataset Selection, Concept Unlearning, Large Language Models, Prompt Selection

# Contents

**Table of Contents** xv

**List of Figures** xvii

**List of Tables** xix

**1 Introduction** 1
- 1.1 Data-Centric AI and the Role of Selection . . . 2
- 1.2 Data Selection for Trustworthy Supervised Learning . . . 2
- 1.3 Trustworthiness Challenges in Large Language Models . . . 3
- 1.4 Prompt Selection for Trustworthy LLM Unlearning . . . 4
- 1.5 A Unified Selection-Centric Perspective . . . 4
- 1.6 Contributions of the Thesis . . . 5
- 1.7 Organization of the Thesis . . . 6

**2 Literature Survey** 7
- 2.1 Introduction . . . 7
- 2.2 Trustworthy AI: Fairness and Robustness . . . 8
- 2.3 Data-Centric AI and Data Valuation . . . 9
- 2.4 Data Subset Selection for Trust Objectives . . . 9
- 2.5 Concept Erasure and Machine Unlearning in LLMs . . . 10
- 2.6 Summary and Research Gaps . . . 11

**3 VTruST: Controllable value function based subset selection for Data-Centric Trustworthy AI** 13
- 3.1 Introduction . . . 13
- 3.2 VTruST: Value-driven Trustworthy AI through Selection of Training Data . 15
  - 3.2.1 A Controllable Value Function-based Framework for DCTAI . . . 15
  - 3.2.2 Value Functions for Trustworthy Data-centric AI . . . 16
  - 3.2.3 An online-OMP algorithm for online sparse approximation . . . 18
- 3.3 Experiments . . . 19
  - 3.3.1 Error rate, Fairness and Robustness on Social Data . . . 19
  - 3.3.2 Accuracy and Robustness on Image and Scientific Datasets . . . 20

3.3.3 Data-centric analysis: Post hoc explanation . . . . . 22
3.4 Conclusions . . . . . 23

**4 MPSelectTune: Prompt-type Selection for Fine-tuning improves Concept Unlearning in LLMs 25**
4.1 Introduction . . . . . 25
4.2 LLM Concept Unlearning . . . . . 27
4.2.1 Problem Definition . . . . . 27
4.2.2 Joint Prediction Prompt . . . . . 29
4.2.3 Loss functions for Concept Unlearning . . . . . 30
4.2.4 Prompt-type selection for LLM Concept Unlearning . . . . . 32
4.3 Experimental Results . . . . . 33
4.3.1 Experimental Setup . . . . . 34
4.3.2 Comparison of Unlearning Performance . . . . . 37
4.3.3 Analysis of Prompts . . . . . 39
4.3.4 Ablation study of loss functions . . . . . 39
4.3.5 Generalization to Unseen Prompt Types . . . . . 40
4.4 Conclusions . . . . . 41

**5 Conclusion and Future Work 43**
5.1 Summary of Contributions . . . . . 43
5.1.1 Controllable Data Subset Selection for Trustworthy AI (VTruST) . 44
5.1.2 Prompt-Type Selection for Robust Concept Unlearning in LLMs (MPSelectTune) . . . . . 45
5.2 Future Work . . . . . 45

**Bibliography 47**

**Publications from the Thesis 57**

# List of Figures

1.1 Overview of the **VTruST** framework. VTruST formulates data subset selection as an optimization over a controllable value function that explicitly balances accuracy, fairness, and robustness objectives to construct trustworthy training datasets. . . . . 3
1.2 Prompt-type selection for robust and worst-case-aware concept unlearning in large language models. . . . . 5

3.1 **Controlling trade-offs in trustworthiness metrics for social data — Adult Census.** . . . . . 21
3.2 **Box plot representation of CF-gap.** . . . . . 23
3.3 **Data Map for randomly taken 5000 samples from TinyImagenet and CIFAR10 augmented training dataset** . . . . . 24
3.4 **Anecdotal samples from VTruST-R & SSR with High Distinctiveness and Uncertainty from TinyImagenet for class Compass.** . . . . . 24

4.1 Overview of the proposed MPSelectTune framework. . . . . 26
4.2 Prompt structure with instruction, exemplars, and test input. . . . . 28
4.3 **Left**: Concept Accuracy for few prompt-types. **Right**: Generalization of the proposed methods to unseen prompt types. . . . . 39

# List of Tables


1.1 Selection mechanisms across learning paradigms. . . . . 5

3.1 **Comparison of VTruST-F with baselines over 60% subset for fairness evaluation.** . . . . . 20
3.2 **Comparison of VTruST-R over varying subset sizes for robustness evaluation. The numbers in brackets indicate the difference with the second best among baselines.** . . . . . 21
3.3 **Performance comparison on scientific datasets** . . . . . 22
3.4 **Sample instances with High Counterfactual Token Fairness Gap** . . . . 23

4.1 Comparison of unlearning performance on BIOS and RT-Gender datasets with LLM-based Baselines. The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics). . . . . 36
4.2 Comparison of unlearning performance on Toxic Bias and Adult Census datasets with LLM-based Baselines. The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics). . . . . 37
4.3 Performance comparison with Pre-LLM baselines (representation unlearning). The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics). . . . . 38
4.4 Unlearning on SciQ-WMDP-Bio Dataset using Llama-2, Llama-3.1, and Mistral-7B . . . . . 38
4.5 Ablation of loss function components in MPSelectTune on Bios Dataset with Llama-2 . . . . . 40

# Chapter 1

# Introduction

Artificial Intelligence (AI) systems have achieved remarkable success across a wide range of applications, including computer vision, natural language processing, recommender systems, and decision support systems. These advances are primarily driven by progress in deep learning architectures [1, 2], the availability of large-scale datasets, and increasing computational resources. Consequently, AI models are now routinely deployed in high-impact real-world domains such as hiring, healthcare, finance, education, and governance.

Despite these successes, the widespread deployment of AI systems has raised serious concerns regarding their trustworthiness. Empirical studies have shown that machine learning models can exhibit unfair or discriminatory behavior due to biased training data [3, 4], fail under distribution shifts and adversarial perturbations [5], and unintentionally memorize or reveal sensitive or harmful information [6, 7]. Such failures undermine user trust and limit the safe adoption of AI systems in sensitive applications.

Trustworthy AI is a multi-faceted concept encompassing fairness, robustness, safety, privacy, and accountability [8, 9]. Importantly, recent evidence suggests that many trust violations are not solely caused by model architecture choices, but are deeply rooted in the data and training signals used to build AI systems. This observation motivates a shift from purely model-centric approaches toward mechanisms that explicitly control training influence.

## 1.1 Data-Centric AI and the Role of Selection

Recent research has advocated a paradigm shift toward *data-centric AI*, where improving data quality, selection, and usage is prioritized over designing increasingly complex models [10, 11]. Several studies have demonstrated that a relatively small subset of training examples can disproportionately influence model predictions, robustness, and bias [12, 13]. This observation highlights the importance of identifying and controlling which data points shape model behavior.

Data subset selection has emerged as an effective approach to address this challenge. Rather than training on the entire dataset, subset selection methods aim to identify informative or influential samples that optimize a given objective. Prior work has explored data selection for computational efficiency [14, 15], robustness [16], and fairness-aware learning [17, 18]. However, most existing methods focus on a single objective and offer limited flexibility in balancing multiple trustworthiness goals simultaneously.

## 1.2 Data Selection for Trustworthy Supervised Learning

In supervised learning, real-world datasets frequently contain redundant, noisy, or biased samples. These imperfections can amplify unfairness, reduce robustness, and lead to unstable generalization under distribution shifts [17, 19]. Traditional mitigation techniques such as reweighting or regularization typically operate implicitly and provide limited interpretability regarding which data points contribute to trust violations.

Data subset selection provides a more explicit and interpretable alternative by directly selecting training samples that optimize specific trust objectives. Figure 1.1 illustrates how data subset selection can act as a control mechanism for trustworthiness in supervised learning.

In the first part of this thesis, we introduce **VTruST**, a value-function-based data subset selection framework for controlling multiple trust objectives, including accuracy, fairness, and robustness. VTruST formulates subset selection as an optimization problem over a

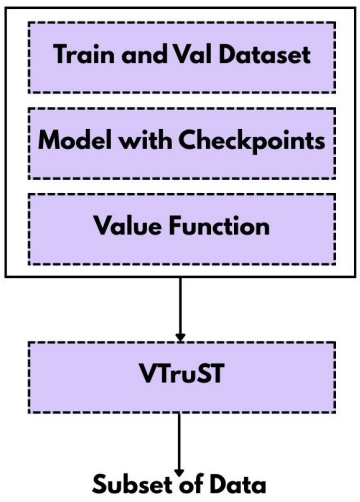



**Figure 1.1** Overview of the **VTruST** framework. VTruST formulates data subset selection as an optimization over a controllable value function that explicitly balances accuracy, fairness, and robustness objectives to construct trustworthy training datasets.

learned value function and employs an efficient online selection strategy inspired by sparse approximation techniques [20, 21]. This framework provides both scalability and interpretability, enabling practitioners to reason about how selected data subsets influence model behavior.

## 1.3 Trustworthiness Challenges in Large Language Models

Large language models (LLMs) such as BERT [22] and LLaMA [23] have demonstrated strong generalization capabilities through large-scale pretraining and fine-tuning. These models underpin a wide range of applications, including conversational agents, content generation, and decision support systems.

Despite their success, LLMs are known to encode sensitive, biased, or harmful concepts that may be revealed depending on how the model is prompted [7, 24]. This prompt-dependent behavior raises serious concerns regarding safety, privacy, and responsible de-

ployment, particularly in open-ended generation settings.

## 1.4 Prompt Selection for Trustworthy LLM Unlearning

Machine unlearning has emerged as a promising approach to mitigate such risks by selectively removing specific knowledge or behaviors from trained models [25, 26]. Existing unlearning methods typically rely on fine-tuning, parameter editing, or representation-level interventions.

Recent studies have shown that unlearning effectiveness is highly prompt-dependent, with certain prompt types inducing significantly higher concept leakage than others [27,28]. Consequently, evaluating unlearning under average-case prompts can obscure worst-case failures.

To address this issue, the second part of this thesis introduces **MPSelectTune**, a prompt-type selection framework that identifies prompt types inducing maximum concept leakage and uses them as adversarial signals during fine-tuning. By explicitly optimizing against worst-case prompt types, MPSelectTune achieves stronger and more reliable unlearning across unseen prompt distributions while preserving task performance (Figure 1.2).

## 1.5 A Unified Selection-Centric Perspective

Although VTruST and MPSelectTune operate in different learning paradigms, they share a common principle: *selection as a control mechanism for trustworthiness*. In supervised learning, selecting influential data points determines the gradients that shape model parameters. In LLM fine-tuning, selecting prompt types determines how internal representations are adapted during learning.

This thesis presents a unified selection-centric view of trustworthy AI, demonstrating that careful selection of training signals can effectively control undesirable behaviors across learning settings.

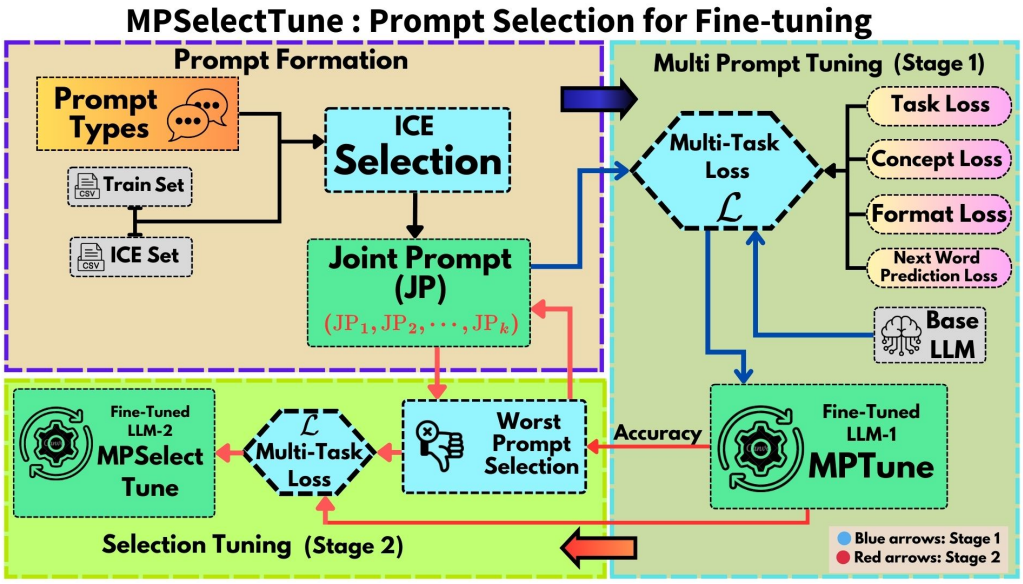



**Figure 1.2** Prompt-type selection for robust and worst-case-aware concept unlearning in large language models.

**Table 1.1** Selection mechanisms across learning paradigms.

| Aspect | Supervised Learning | LLM Fine-Tuning |
|---|---|---|
| Selection Unit | Data samples | Prompt types |
| Trust Objective | Fairness, robustness | Safety, unlearning |
| Failure Mode | Biased gradients | Concept leakage |

## 1.6 Contributions of the Thesis

The main contributions of this thesis are summarized as follows:

- We propose **VTruST**, a controllable data subset selection framework for optimizing multiple trustworthiness objectives in supervised learning.
- We introduce **MPSelectTune**, a prompt-type selection strategy for robust and worst-case-aware concept unlearning in large language models.
- We present a unified selection-centric perspective that connects data-centric AI and LLM unlearning under a common trust control framework.

## 1.7 Organization of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2** reviews related work on data-centric AI, data subset selection, trustworthy machine learning, and machine unlearning.
- **Chapter 3** presents the proposed VTruST framework for trustworthy supervised learning.
- **Chapter 4** introduces MPSelectTune for prompt-type selection in LLM unlearning.
- **Chapter 5** concludes the thesis and discusses directions for future research.

# Chapter 2

# Literature Survey

## 2.1 Introduction

As discussed in Chapter 1, the central theme of this thesis is to improve the trustworthiness of AI systems by explicitly controlling the training signals that shape model behavior. Specifically, this thesis investigates two complementary settings: (i) supervised learning, where training data selection plays a critical role in determining fairness and robustness, and (ii) large language model (LLM) fine-tuning, where prompt selection strongly influences safety and concept leakage.

In this chapter, we present a comprehensive review of prior work related to these themes. We first survey literature on trustworthy AI, highlighting fairness, robustness, and their inherent trade-offs. We then review data-centric AI and data subset selection methods, which motivate selection-based control mechanisms in supervised learning. Next, we discuss existing approaches to machine unlearning and concept erasure, with a focus on challenges unique to LLMs. Finally, we summarize the limitations of existing methods that motivate the approaches proposed in this thesis.

The chapter is organized as follows:

- Section 2.2 reviews literature on trustworthy AI, including fairness and robustness.

- Section 2.3 surveys data-centric AI and data valuation methods.
- Section 2.4 discusses data subset selection for trust objectives.
- Section 2.5 reviews concept erasure and machine unlearning in LLMs.
- Section 2.6 summarizes research gaps and positions this thesis.

## 2.2 Trustworthy AI: Fairness and Robustness

Trustworthy AI aims to ensure that machine learning systems behave reliably, ethically, and safely when deployed in real-world settings [11]. Among the many dimensions of trustworthiness, fairness and robustness have received the most attention.

Fairness-focused research seeks to mitigate discriminatory outcomes across sensitive attributes such as gender or race. Early approaches include representation learning methods that enforce group-level parity [29], reweighting or resampling strategies [18, 30], and data augmentation techniques [31, 32]. Robustness-oriented research, on the other hand, focuses on improving model stability under distribution shifts or adversarial perturbations [16].

Recent studies have highlighted inherent trade-offs between trustworthiness objectives. For example, improving fairness can reduce accuracy [18], while enhancing robustness may degrade standard performance [33]. Other works analyze the interaction between fairness and robustness [17]. These trade-offs indicate that no single model-centric solution can simultaneously optimize all trust objectives.

Most existing methods operate by modifying model architectures, training objectives, or optimization constraints. While effective, such approaches offer limited interpretability and do not provide explicit control over how different trust objectives are balanced.

## 2.3 Data-Centric AI and Data Valuation

Data-centric AI (DCAI) emphasizes improving model performance and trustworthiness by optimizing data rather than model architecture [10, 11]. A growing body of work shows that training data quality and composition significantly influence model behavior.

Several studies focus on identifying easy, hard, or ambiguous samples using confidence-based heuristics [34, 35] or loss dynamics [36]. Other works aim to detect mislabeled or low-quality samples [37].

Another prominent line of research explores data valuation, where each training sample is assigned an importance score. Methods include Shapley-value-based approaches [13, 38], influence functions [12, 39], reinforcement learning [40], gradient-based approximations [15, 41], and training-free metrics [42, 43]. These techniques primarily optimize for predictive performance and often incur high computational costs.

Importantly, most existing data valuation methods focus solely on accuracy and do not explicitly account for fairness or robustness. Moreover, they do not provide a user-controllable mechanism to balance multiple trust objectives. Recent work by Roh et al. [17] and Liang et al. [11] highlight this gap between data valuation and trustworthiness.

## 2.4 Data Subset Selection for Trust Objectives

Data subset selection methods aim to identify a subset of training samples that preserves or improves model performance while reducing computational or trust-related costs. Coreset methods select representative samples for efficiency [14], while gradient-based methods select influential points [15].

Some recent works explicitly consider trust objectives. Roh et al. [17] propose selecting samples to improve both fairness and robustness by balancing sensitive groups. While effective, such methods rely on fixed constraints and do not allow continuous trade-offs between objectives.

Overall, existing approaches lack a unified framework that allows explicit, user-controllable trade-offs across multiple trustworthiness dimensions. This limitation motivates the need for flexible, value-function-based data selection mechanisms.

## 2.5 Concept Erasure and Machine Unlearning in LLMs

Machine unlearning aims to remove specific information or behaviors from trained models without full retraining. Early work on concept erasure focused on removing linear correlations between representations and sensitive attributes [44]. Subsequent methods extended this to kernelized and non-linear settings [45, 46].

With the advent of large language models (LLMs), unlearning has become more challenging due to large-scale pretraining and entangled representations. Recent works study information unlearning for privacy and safety [24, 47]. Gradient ascent-based fine-tuning [6, 26], dememorization [48, 49], and parameter editing [50] have shown promise.

Another line of work explores in-context or post-hoc interventions, such as prompt-based filtering [27] or external negation models [25]. However, these methods often depend on carefully designed prompts and may fail under adversarial or unseen prompt distributions.

Recent studies reveal that concept leakage in LLMs is highly prompt-dependent, and evaluating unlearning under average-case prompts can mask worst-case failures [28]. Existing methods largely overlook this prompt sensitivity.

**RL-based Safety Methods.** Reinforcement learning from human feedback (RLHF) [51] and Direct Preference Optimization (DPO) [52] have been used for safety alignment by training reward models that penalize harmful outputs. While effective for broad alignment, these methods require expensive reward model training and human annotation. In contrast, MPSelectTune (Chapter 4) operates via supervised fine-tuning with adversarial prompt selection, avoiding reward model dependencies while achieving targeted concept removal.

## 2.6 Summary and Research Gaps

The literature reveals several important gaps. First, most trustworthy AI methods are model-centric and offer limited interpretability or control. Second, data valuation and subset selection methods primarily optimize accuracy and lack mechanisms to balance fairness and robustness. Third, existing unlearning approaches for LLMs do not explicitly account for prompt-dependent worst-case behavior.

This thesis addresses these gaps by proposing selection-based control mechanisms in both supervised learning and LLM fine-tuning. By framing trustworthiness as a selection problem, over data samples or prompt types, the proposed approaches offer scalable, interpretable, and controllable solutions for trustworthy AI.

# Chapter 3

# VTruST: Controllable value function based subset selection for Data-Centric Trustworthy AI

## 3.1 Introduction

Trustworthiness [8, 9] of predictions made by Machine Learning models is crucial in many applications. In applications impacting society, e.g. loan eligibility prediction [3], criminal recidivism risk prediction [53], etc, fairness in prediction across different marginalized groups is as important as overall prediction accuracy. Similarly, the robustness of object detection systems for autonomous driving against perturbed input images [54], or robustness against label corruption in phase transition prediction of sub-atomic particles [55] are important metrics compared to overall prediction accuracy. Tradeoffs between various notions of fairness with accuracy, e.g. individual fairness (demographic parity/equalized odds) [18, 30] or group fairness (Accurate Fairness) [56] are being studied for different models and training mechanisms. Similar studies have also been reported on inherent tradeoffs between feature robustness in images [57, 58] with accuracy, and adversarial label robustness [33, 59]. While inherent tradeoffs between the trustworthiness metrics such as accuracy vs fairness or robustness is generally accepted, the nature of additional bias in-

troduced by algorithms e.g adversarial training [59] or FairBatch [18] is not clear. In this work, we follow a data-centric approach to designing trustworthy AI techniques.

Data-centric approaches to AI model development [10] strive to design methods for creating high-quality training datasets, that when used with standard SGD-based training algorithm can lead to models with specific trustworthiness properties. This eliminates the algorithmic bias introduced by specific algorithms, while limiting the bias only to the newly created training dataset, which is easier to interpret. While many data valuation [12, 13, 39] techniques have been developed for the selection of high-quality training data subsets, most of them optimize only one property, e.g. validation set error rate. While well-known metrics for robustness and fairness exist, their use as a viable and efficient value function remains to be studied. A key research issue in designing a data-centric approach is the design of an appropriate "value function" that captures the notion of value of a training datapoint (toward trustworthiness metrics) while also being efficiently optimizable. Another important research question is: can the value functions corresponding to various trustworthiness metrics be combined into a single value function using user-defined weightage? In this work we address these research questions, effectively leading to a general data-centric framework to achieve user-controlled tradeoffs between different trustworthiness metrics.

We propose additive value functions for accuracy, fairness, and robustness, which can be combined to form composite value functions. The additiveness of the value functions is a key property that allows us to pose the problem of training data valuation and selection as an *online sparse approximation problem*. We propose a novel *online orthogonal matching pursuit (OMP)* algorithm that greedily replaces features corresponding to a selected datapoint with those of a new datapoint, if there is a net improvement in the overall value of the selected set. Unlike the traditional OMP [60] which makes a pass through the entire training dataset to select an example, the proposed online OMP makes a pass through the selected datapoints (a much smaller set) at the time of training update to optionally replace an existing selected point. Experimental results on various applications demonstrate that models trained on subsets selected by VTruST can outperform all state-of-the-art baselines by $\sim 10 - 20\%$ and can also provide data-centric explanations behind its performance.

## 3.2 VTruST: Value-driven Trustworthy AI through Selection of Training Data

We propose a controllable value function-based framework for developing trustworthy models using a data-centric paradigm. Our system has two components: (1) A general value function-based framework that allows users to specify a combination of trustworthiness metrics (sections 3.2.1 and 3.2.2), and (2) a novel online subset selection algorithm for constructing high-quality training dataset based on the specified value function (section 3.2.3).

### 3.2.1 A Controllable Value Function-based Framework for DCTAI

Let $\mathcal{D} = \{d_i | i = 1, 2, .., N\}$ be the training dataset and $\mathcal{D}' = \{d'_j | j = 1, 2, .., M\}$ be the validation dataset. Every datapoint $d' \in \mathcal{D}'$ can be used to define the value function $\mathcal{V}(\theta, d')$ which is used for calculating the value of a model $\theta$. Given a run of model training, we define the incremental value function $v_i^t(d'_j)$ as the decrease in loss incurred due to an SGD update [61] using the datapoint $d_i$: $v_i^t(d'_j) = l(\theta_t^{i-1}, d'_j) - l(\theta_t^i, d'_j)$, where $\theta_t^{i-1}$ and $\theta_t^i$ are the model parameters before and after the SGD update involving the training datapoint $d_i$ in the $t^{th}$ epoch. Hence the value of a model $\theta^T$ can be defined as: $\mathcal{V}(d'_j) = \sum_{t=1}^{T} \sum_{i=1}^{N} v_i^t(d'_j)\ \forall d' \in \mathcal{D}'$. We overload the notation to define the value function vector $\mathcal{V}(\mathcal{D}') = \sum_{t=1}^{T} \sum_{i=1}^{N} v_i^t(\mathcal{D}')$, where $v_i^t(\mathcal{D}') \in \mathbb{R}^M$ is the vector of incremental values over all validation set datapoints.

Our data-centric framework aims to find a subset of training datapoints $\mathcal{S} \in \mathcal{D}$ that leads to a high-value model $\theta^t$ after training for $t$-epochs. Let $\vec{y}_t = \sum_{k=1}^{t} \sum_{i=1}^{N} v_i^k(\mathcal{D}')$ be the cumulative value function till the $t^{th}$ epoch. We formulate the training data subset selection problem as a sparse approximation: $\vec{y}_t \approx \sum_{d_i \in S \subseteq \mathcal{D}} \alpha_i [\sum_{k=1}^{t} v_i^k(\mathcal{D}')]$ , where $\alpha_i$ are the weights for the selected training datapoint $d_i$. Next, using a second order Taylor series expansion of the change in loss function and plugging in the SGD update $\theta_t^i - \theta_t^{i-1} = \eta_t \nabla l(\theta_t^{i-1}, d_i)$, we obtain the following approximation for each term in the value function $l(\theta_t^i, \mathcal{D}') - l(\theta_t^{i-1}, \mathcal{D}') \approx \eta_t \nabla l(\theta_t^{i-1}, d_i)^T \nabla l(\theta_t^{i-1}, \mathcal{D}') + \mathcal{O}(||\theta_t^i - \theta_t^{i-1}||_2^2)$.

We truncate the Taylor expansion till the second-order terms to arrive at the following sparse approximation problem: $\vec{y}_t \approx \sum_{d_i \in S} \alpha_i \left[\sum_{k=1}^{t} \vec{X}_i^k\right] \ \forall t = 1, ..., T$ where $\vec{X}_i^k = \nabla l(\theta_k^{i-1}, d_i)^T \nabla l(\theta_k^{i-1}, \mathcal{D}') + \frac{(\nabla l(\theta_k^{i-1}, d_i)^T \nabla l(\theta_k^{i-1}, \mathcal{D}'))^2}{2}$ are the features for the $i^{th}$ training point calculated in epoch $t$. We use $\vec{y}_t$ and $\vec{X}_i^t$ to denote the predictor and predicted variables for valuating training datapoint $d_i$ using the entire validation set $\mathcal{D}'$. The main challenge in solving this approximation problem is that we need to store all the features $\vec{X}_i^k$ for all training datapoints $i$ over epochs $k = 1, ..., t$, in order to compute $\sum_{k=1}^{t} \vec{X}_i^k$. This becomes prohibitively expensive. Instead, we solve the following *online sparse approximation* (OSA) problem for each epoch $t$:

$$\vec{y}_t \approx \sum_{(p,q) \in S_t} \beta_p^q \vec{X}_p^q \tag{3.1}$$

Here, $S_t$ is the set of selected training datapoints after epoch $t$. Note that the set $S_t$ can contain datapoints indexed by $p$ with features from any of the epochs $q = 1, ..., t$. We constrain the size of $S_t$ to be less than a user-specified parameter $\omega$. We describe an online algorithm for solving the above problem in Section 3.2.3. Note that the value function $\mathcal{V}(\mathcal{D}')$ only needs to be additive over the training datapoints and epochs for the above formulation to be valid. Hence, this framework applies to a composite value function $\mathcal{V}(\mathcal{D}') = \sum_f \lambda_f \mathcal{V}_f(\mathcal{D}')$, where each value function $\mathcal{V}_f(.)$ satisfies the additive property. This leads us to a general *controllable* framework for incorporating many trustworthiness value functions, controlled using the user-specified weights $\lambda_f$.

### 3.2.2 Value Functions for Trustworthy Data-centric AI

For the accuracy metric, we use the value function proposed in [61], which is defined as the decrease in loss incurred due to an SGD update using the datapoint $d_i$: $v_i^t(d'_j) = l(\theta_t^{i-1}, d'_j) - l(\theta_t^i, d'_j)$ where $\theta_t^{i-1}$ and $\theta_t^i$ are the model parameters before and after the SGD update involving the training datapoint $d_i$ in the $t^{th}$ epoch. Hence, the **accuracy value function** vector in defined as $\mathcal{V}_a(\mathcal{D}') = \sum_{t=1}^{T} \sum_{i=1}^{N} v_i^t(\mathcal{D}')$.

**Robustness Value Function**: Training data augmentation using various perturbations has been observed to improve robust accuracy [62, 63]. We use various perturbations to create

the augmented training set $\mathcal{D}_a$ and validation set $\mathcal{D}'_a$ from $\mathcal{D}$ and $\mathcal{D}'$ respectively. The robustness value function is defined as $\mathcal{V}_r(\mathcal{D}'_a) = \sum_{t=1}^{T} \sum_{d_i \in \{\mathcal{D} \cup \mathcal{D}_a\}} l(\theta_t^i, \mathcal{D}'_a) - l(\theta_t^{i-1}, \mathcal{D}'_a)$. Since $\mathcal{V}_r$ is derived from the loss function (that is additive), it also follows the additive property.

**Fairness Value Function**: Existing literature in fairness [18, 30] uses equalized odds (EO) disparity and demographic parity disparity for achieving fair models. Let $x \in \mathcal{X}$ be the input domain, $\{y_0, y_1\} \in \mathcal{Y}$ be the true binary labels, and $\{z_0, z_1\} \in \mathcal{Z}$ be the sensitive binary attributes. We define the fairness value function as the change in EO disparity: $\mathcal{V}_f(\mathcal{D}') = \sum_{t=1}^{T} \sum_{d_i \in \mathcal{D}} ed(\theta_t^i, \mathcal{D}') - ed(\theta_t^{i-1}, \mathcal{D}')$. Based on [64], it is defined as the maximum difference in accuracy between the sensitive groups ($z \in \mathcal{Z}$) pre-conditioned on the true label ($y \in \mathcal{Y}$): $ed(\theta, \mathcal{D}') = max(\|l(\theta, \mathcal{D}'_{y_0,z_0}) - l(\theta, \mathcal{D}'_{y_0,z_1})\|, \|l(\theta, \mathcal{D}'_{y_1,z_0}) - l(\theta, \mathcal{D}'_{y_1,z_1})\|)$. Considering we have $ed(\theta, \mathcal{D}')$ defined for two validation sets $\mathcal{D}'_1$ and $\mathcal{D}'_2$ and the loss function is inherently additive, $ed(\theta, \mathcal{D}'_1) + ed(\theta, \mathcal{D}'_2) = ed(\theta, (\mathcal{D}'_1 + \mathcal{D}'_2))$ also holds true, thus $\mathcal{V}_f$ turning out to be additive.

**Composite value functions:** We can combine the value functions for accuracy ($\mathcal{V}_a(\mathcal{D}')$), robustness ($\mathcal{V}_r(\mathcal{D}'_a)$) and fairness ($\mathcal{V}_f(\mathcal{D}')$) to construct different composite value functions for observing tradeoffs between different trustworthiness metrics. We use the following combinations for our experiments: (a) Accuracy-Fairness : $\mathcal{V}_{af}(\mathcal{D}') = \lambda\mathcal{V}_a(\mathcal{D}') + (1 - \lambda)\mathcal{V}_f(\mathcal{D}')$ ; (b) Accuracy-Robustness : $\mathcal{V}_{ar}(\mathcal{D}', \mathcal{D}'_a) = \lambda\mathcal{V}_a(\mathcal{D}') + (1 - \lambda)\mathcal{V}_r(\mathcal{D}'_a)$ ; (c) Robustness-Fairness : $\mathcal{V}_{rf}(\mathcal{D}', \mathcal{D}'_a) = \lambda\mathcal{V}_r(\mathcal{D}'_a) + (1 - \lambda)\mathcal{V}_f(\mathcal{D}')$. The user-defined parameter $\lambda$ is used to control the tradeoff between the two objectives.

### 3.2.3 An online-OMP algorithm for online sparse approximation

**Algorithm 1:** : **VTruST**

1: **Input:**
i. $\omega$ : Total number of datapoints to be selected
ii. $\vec{y}$ : Targeted value function
iii. $\vec{X}_i$ : Features of all training points $d_i \in \mathcal{D}$
iv. $S$ : Set of selected datapoint indices
v. $\vec{\beta} \in \mathbb{R}^{|S|}$: Weight of selected datapoints
2: **Initialize:**
$S \longleftarrow \phi$ //Indices of selected datapoints
3: **for** each epoch $t \in \{1, 2, \ldots, T\}$ **do**
4: **for** each datapoint $d_i \in \mathcal{D}$ **do**
5: **Input:** $\vec{y}_t, X_i^t \quad \forall i \in \{1, 2, .., N\}, ||X_i^t||_2 = 1$
6: **Process:**
7: **if** $|S_{t-1}| = \omega$ **then**
8: $S_t \leftarrow$ **DataReplace**($\vec{y}_t, \vec{\xi}_{t-1}, S_{t-1}, \vec{\beta}_{t-1}, \vec{X}_i^t$)
9: **else**
10: $S_t \longleftarrow S_{t-1} \cup \{i\}$
// Add datapoints till the cardinality of $S_t$ reaches $\omega$
11: **end if**
12: Update $\vec{\beta}_t = \text{argmin}_\beta \|\vec{y}_t - \sum_{p,q \in S_t} (\beta_q^p \vec{X}_q^p)||_2$
13: Update $\vec{\xi}_t = \sum_{p,q \in S_t} \beta_q^p \vec{X}_q^p$
14: **end for**
15: **end for**
16: **Output**:Final
set of selected datapoint indices $S_T$, learned coefficients $\{\beta_q^p | p, q \in S_T\}$

**Algorithm 2:** :**DataReplace**($\vec{y}_t, \vec{\xi}_{t-1}, S_{t-1}, \vec{\beta}_{t-1}, \vec{X}_i^t$) - *Replace an existing datapoint.*

1: $\vec{\rho}_t = \vec{y}_t - \vec{\xi}_{t-1}$
2: $\pi_{max}$ = -$\infty$
3: $(a, b) = \phi$
4: $\pi \longleftarrow \text{abs}(\vec{X}_i^t . \vec{\rho}_t)$
5: **for** each index $p, q \in S_{t-1}$ **do**
6: $\pi' \longleftarrow \text{abs}(\vec{X}_q^p . \vec{\rho}_t)$
7: $\gamma \longleftarrow \beta_q^p$
8: **if** $\pi > \pi'$ & $\gamma \leq 0$ & $(\pi' + \gamma) > \pi_{max}$ **then**
9: $\pi_{max} \longleftarrow \pi' + \gamma$
10: $a, b \longleftarrow p, q$
11: **end if**
12: **end for**
13: **if** $(a, b) \neq \phi$ **then**
14: $S_t \longleftarrow S_{t-1} \setminus \{a, b\} \cup \{t, i\}$
15: **end if**
16: return $S_t$

In this section, we describe a novel online-OMP-based algorithm for the *online sparse approximation problem*(OSA) in Algorithm 1. The key difference between OSA and standard sparse approximation setting is that in OSA, new columns $\vec{X_p^q}$ are added and the target value $\vec{y_t}$ is updated at each epoch $t$. Line 10 in Algorithm 1 adds new datapoints till the cardinality of $S_t$ reaches $\omega$. Once the buffer is saturated, the *DataReplace* module is invoked in line 8 to replace an existing selected datapoint with the incoming datapoint. The criteria for replacement is to select the datapoints in $S_t$ that contribute to a better approximation of the current value function $\vec{y}_t$. Hence a new datapoint with features $\vec{X}_i^t$ gets selected if the current approximation error reduces after the replacement. We compute the projection of the incoming datapoint features, $\vec{X}_i^t$ and that of the features of the selected datapoints $\vec{X}_q^p \; \forall p, q \in S_t$ on the existing residual vector $\vec{\rho}_t$, measured by $\pi$ and $\pi'$ respectively. We also denote by $\gamma$, the contribution of datapoint $p, q \in S_t$ obtained through $\beta_q^p$. The datapoints with indices $(p, q)$ in $S_t$ whose additive impact $(\pi' + \gamma)$ is smaller than that of incoming datapoint $(i, t)$, but larger than the current feature for replacement $(\vec{X}_q^p)$ (line 8), gets substituted with the incoming point in line 14 of Algorithm 2. In terms of complexity, the per-epoch time complexity of OMP is $\mathcal{O}(\omega MN)$ and that of *VTruST* is $\mathcal{O}(\omega M(N-\omega))$.

**Hyperparameter selection**: The proposed framework has two user-controlled hyperparameters, the tradeoff $\lambda$ and the subset-size $\omega$. Since the metrics are not monotone in $\omega$, we perform a grid search with various selection fractions between 10 - 90%. Exploiting the monotonicity of metrics w.r.t. $\lambda$, one can fix a threshold on the first metric, say accuracy in case of $\mathcal{V}_{af}$ and perform a binary search to arrive at an optimal point w.r.t. the second metric (fairness in $\mathcal{V}_{af}$), once the threshold w.r.t. the first metric has been satisfied.

## 3.3 Experiments

In this section, we describe the datasets, models, and evaluation metrics used for the trustworthiness metrics - Accuracy, Fairness and Robustness.

We analyze the performance of *VTruST* (VTruST-F with $V_{af}$, VTruST-R with $V_{ar}$, VTruST-FR with $V_{rf}$) over various applications. All our experiments have been executed on a single Tesla V100 GPU.

### 3.3.1 Error rate, Fairness and Robustness on Social Data

We evaluate the ability of VTruST to achieve a tradeoff between pairs of the three important social trustworthiness metrics: error rate (ER), fairness and robustness. Our *baselines* are: Wholedata standard training (ST), Random, SSFR [17], FairMixup [65] and FairDummies [30]. We report results on three benchmark datasets: COMPAS [53] , Adult Census [66] and MEPS-20 [67].

We use a 2-layer neural network for all the datasets. We report two fairness metrics: equalised odds (EO) Disparity [3] and demographic parity(DP) Disparity [4] following [17].

**Fairness and Error Rate comparison (VTruST-F) with baselines**: We compare the performance metrics of VTruST-F with the baselines in Table 3.1.

The better the model is, the lower its ER as well as its fairness measures. We can

observe in Table 3.1 that VTruST-F with 60% selected subset outperforms all the other methods in terms of fairness measures by a margin of $\sim 0.01 - 0.10$, and performs close to Wholedata-ST that yields the lowest ER. This denotes that it is able to condemn the error-fairness tradeoff emerging out to be the best performing method. We report these results with standard deviation across 3 runs.

**Table 3.1 Comparison of VTruST-F with baselines over 60% subset for fairness evaluation.**

| Methods | COMPAS | | | AdultCensus | | | MEPS20 | | |
|---|---|---|---|---|---|---|---|---|---|
| | ER ±std | EO Disp ±std | DP Disp ±std | ER ±std | EO Disp ±std | DP Disp ±std | ER ±std | EO Disp ±std | DP Disp ±std |
| Wholedata-ST | 0.34 ±0.001 | 0.31 ±0.05 | 0.24 ±0.03 | 0.16 ±0.002 | 0.19 ±0.06 | 0.13 ±0.06 | 0.09 ± 0.001 | 0.09 ± 0.007 | 0.08 ± 0.0008 |
| Random | 0.35 ±0.002 | 0.20 ±0.10 | 0.23 ±0.09 | 0.19 ±0.002 | 0.16 ±0.05 | 0.13 ±0.05 | 0.12 ±0.017 | 0.06 ±0.02 | 0.08 ±0.005 |
| SSFR | 0.35 ±0.002 | 0.26 ±0.03 | 0.17 ±0.02 | 0.21 ±0.001 | 0.18 ±0.03 | 0.12 ±0.01 | 0.14 ±0.003 | 0.10 ±0.011 | 0.06 ±0.005 |
| Fair-Dummies | 0.35 ±0.002 | 0.24 ±0.02 | 0.17 ±0.01 | 0.16 ±0.002 | 0.14 ±0.01 | 0.10 ±0.01 | 0.12 ±0.001 | 0.13 ±0.005 | 0.08 ±0.003 |
| Fair-Mixup | 0.35 ±0.03 | 0.15 ±0.03 | 0.13 ±0.04 | 0.24 ±0.04 | 0.11 ±0.05 | 0.1 ±0.02 | 0.89 ±0.02 | 0.02 ±0.04 | 0.05 ±0.03 |
| VTruST-F | 0.34 ±0.002 | **0.15 ±0.01** | **0.13 ±0.01** | 0.18 ±0.001 | **0.11 ±0.03** | **0.05 ±0.01** | 0.09 ±0.003 | **0.01 ±0.001** | **0.05 ±0.0008** |

**Tradeoffs between Error rate, Fairness and Robustness (VTrust-F , VTruST-FR)**: We observe the tradeoffs between error rate vs fairness (VTruST-F: Figure 1a) and fairness vs robustness (VTruST-FR: Figure 1b) through pareto frontal curve by varying $\lambda \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$.

The error rate is measured on the clean test sets while robust error rates are measured on the label flipped test sets.

We can observe that Wholedata-ST has a lower error rate but high disparity and robust error values. The other baselines continue to have a higher error rate and disparity compared to VTruST.

### 3.3.2 Accuracy and Robustness on Image and Scientific Datasets

We evaluate VTruST on three image datasets: CIFAR10 [68] , MNIST [69] and Tiny-imagenet [70] using ResNet-18 [71]. For evaluation, we use the standard accuracy (SA) computed on the clean test sets and the robust accuracy (RA) computed on the corrupted

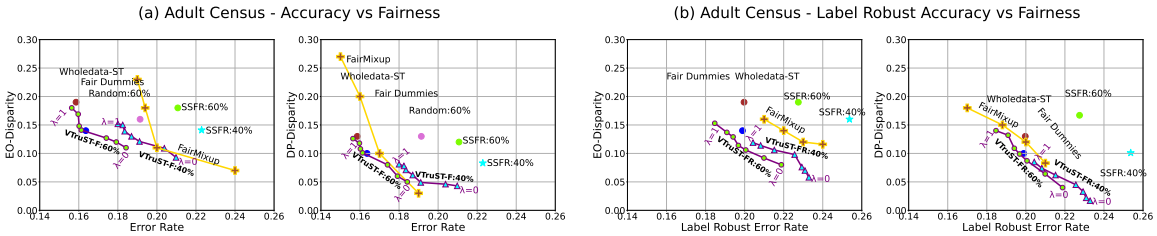



**Figure 3.1 Controlling trade-offs in trustworthiness metrics for social data — Adult Census.**

**Table 3.2 Comparison of VTruST-R over varying subset sizes for robustness evaluation. The numbers in brackets indicate the difference with the second best among baselines.**

| Methods | MNIST | | | CIFAR10 | | | TinyImagenet | | |
|---|---|---|---|---|---|---|---|---|---|
| | #Data points | SA | RA | #Data points | SA | RA | #Data points | SA | RA |
| Clean-ST | 60K | 99.35 | 87.00 | 50K | 95.64 | 83.95 | 100K | 63.98 | 23.36 |
| AugMax | 240K | 97.62 | 88.79 | 200K | 94.74 | 86.44 | 400K | 54.82 | 40.98 |
| SAug | 260K | **99.36** (1.74) | **97.31** (8.52) | 200K | **94.9** (0.16) | **90.13** (3.69) | 300K | **62.04** (7.22) | **42.04** (1.06) |
| **After subset selection from SAug** | | | | | | | | | |
| SSR:40% | 104K | 98.98 | 94.96 | 80K | 93.3 | 85.73 | 120K | 32.82 | 24.42 |
| VTruST-R:40% | 104K | **99.04** (0.06) | **96.29** (1.33) | 80K | **94.74** | **88.23** (1.79) | 120K | **57.3** (2.48) | 39.69 |
| SSR:60% | 156K | 99.07 | 96.53 | 120K | 93.77 | 88.0 | 180K | 41.94 | 30.07 |
| VTruST-R:60% | 156K | **99.12** (0.05) | **97.09** (0.56) | 120K | **94.77** (0.03) | **89.21** (1.21) | 180K | **60.88** (6.03) | **41.50** (0.52) |

test sets, CIFAR-10-C, Tiny ImageNet-C [5] and MNIST-C [72]. While augmentation leads to robustness [62], it also leads to a large dataset with redundancy. We use VTruST-R to select high-quality subset from augmented data.

Empirically we find that creation of augmented data by sampling images based on how difficult an augmentation is leads to better performance compared to uniform selection across augmentations. Next, we define the baselines. (i) *Clean-ST*: Unaugmented training dataset. (ii) *Uniform Augmentation (UAug)* (iii) *Sampled Augmentation (SAug)* (iv) *SSR* [17]: Training subset using the robustness objective function. (v) *AugMax* [16].

**Robustness and Accuracy comparison (VTruST-R) with baselines:** We compare VTruST-R with the baselines in Table 3.2 where it can be seen that model trained on clean datasets (*Clean-ST*) performs abysmally in terms of RA, indicating the need of data augmentations. VTruST-R is seen to outperform AugMax in most of the scenarios, thus indicating that data-centric approaches help in creating quality training datasets.

**Table 3.3 Performance comparison on scientific datasets**

| Metrics | Spinodal | | | | | | | EOSL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Whole data | Rand 40% | SSFR 40% | VTruST -R 40% | Random 60% | SSFR 60% | VTruST -R 60% | Whole data | Rand 40% | SSFR 40% | VTruST -R 40% |
| SA | **83.08** | 73.05 | 74.84 | **80.33** | 77.06 | 78.94 | **81.93** | 70.01 | 63.74 | 62.40 | **66.10** |
| RA | 76.89 | 61.11 | 62.32 | **78.36** | 75.11 | 75.18 | **80.41** | 66.72 | 60.04 | 56.90 | **65.27** |

**Scientific datasets** : We analyzed the performance of VTruST-R on binary class scientific datasets - Spinodal and EOSL [55] that have 29,000 samples with 400 features and 180,000 samples with 576 features respectively. We used the experimental setup as [55] for evaluation. Table 3.3 shows that VTruST-R (using label flipping for robustness) performs close to the wholedata in standard accuracy (SA) and better in terms of robust accuracy (RA).

### 3.3.3 Data-centric analysis: Post hoc explanation

In this section, we explore the characteristics of the selected samples to justify their quality.

**Explanation for fairness:**

We use the metric *Counterfactual Token Fairness Gap (CF-Gap)* [73] for our evaluation. Given a selected instance $x$, we generate a counterfactual instance $x'$ by altering its sensitive attribute and define CF-Gap$(x)$ as $\|f(x_i) - f(x_i')|$ where $f(x)$ corresponds to the model confidence on the target label.

We plot the distribution of CF-Gap in Figure 3.2. It can be observed that VTruST-F acquires the least value, justifying its retainment of fair subsets leading to fair models.

We show 10 anecdotal samples from the Adult Census dataset in Table 3.4 on the basis of high *CF-Gap* and we can observe that SSFR has a large number of redundant samples with similar attribute values (highlighted) while VTruST-F which anyway has relatively lower CF-gap contains a diverse set of samples.

**Figure 3.2 Box plot representation of CF-gap.**

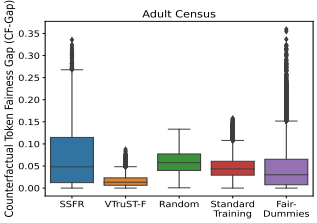



**Table 3.4 Sample instances with High Counterfactual Token Fairness Gap**

| VTruST-F | | | | |
|---|---|---|---|---|
| *Feat* | Rel | Race | Sex | NC |
| $D_1$ | ORel | B | F | JM |
| $D_2$ | NIF | W | M | US |
| $D_3$ | NIF | W | M | US |
| $D_4$ | OC | API | F | TW |
| $D_5$ | Husb | W | M | US |
| $D_6$ | UnM | W | F | US |
| $D_7$ | Wife | W | F | US |
| $D_8$ | OC | W | M | US |
| $D_9$ | NIF | AIE | F | Col |
| $D_{10}$ | UnM | W | F | DE |

| SSFR | | | | |
|---|---|---|---|---|
| *Feat* | Rel | Race | Sex | NC |
| $D_1$ | Husb | W | M | US |
| $D_2$ | Husb | W | M | US |
| $D_3$ | Husb | W | M | US |
| $D_4$ | Husb | W | M | US |
| $D_5$ | Husb | W | M | US |
| $D_6$ | Husb | W | M | US |
| $D_7$ | NIF | W | M | US |
| $D_8$ | OC | W | M | US |
| $D_9$ | Husb | W | M | DE |
| $D_{10}$ | OC | W | M | US |

**Explanation for robustness:** Delving into the literature [34,74], we pick two measures - *uncertainty* and *distinctiveness*. Having a set of hard-to-learn and distinguishable samples in the subsets makes the model more generalizable and robust.

We quantify uncertainty of an instance $x$ as predictive entropy $(-f(x)logf(x))$ and distinctiveness as $\mathbb{E}_{e\in\mathcal{S}}dist(fv(x), fv(e))$ where $dist(,)$ is the euclidean distance and $fv(.)$ is the feature from the model's penultimate layer. Based on the data maps in Figure 3.3, we show anecdotal samples in Figure 3.4 having High Distinctiveness-High Uncertainty (HD-HU). The anecdotal samples and the histogram visualization show that VTruST-R selects diverse samples with difficult augmentations like impulse noise and glass blur, while similar(mostly white-background) and no-noise or easier augmentation-based samples like brightness are more observed in SSR samples, thus justifying the robust selection across augmentations using VTruST-R.

## 3.4 Conclusions

In summary, existing approaches to trustworthy AI have primarily focused on designing fair or robust models through model-centric interventions, while a separate line of work has explored data valuation techniques largely optimized for predictive accuracy. Although some recent methods investigate trade-offs between trustworthiness objectives such as fairness, robustness, and accuracy, they typically rely on fixed constraints and lack mechanisms for explicit, user-controllable trade-offs. Moreover, most data-centric AI

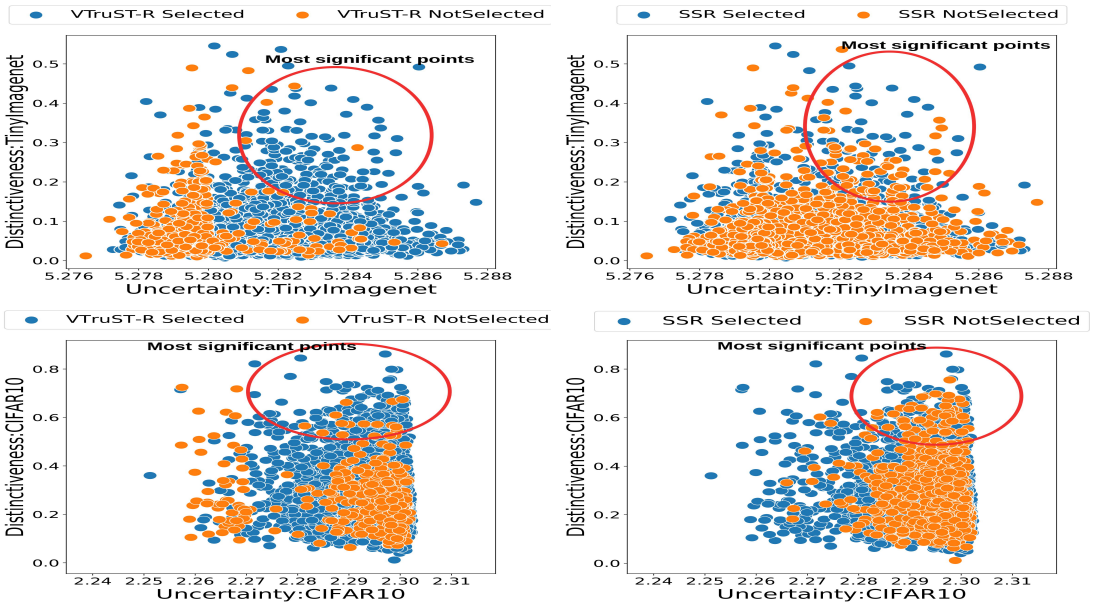



**Figure 3.3 Data Map for randomly taken 5000 samples from TinyImagenet and CIFAR10 augmented training dataset**


**Figure 3.4 Anecdotal samples from VTruST-R & SSR with High Distinctiveness and Uncertainty from TinyImagenet for class Compass.**

methods do not incorporate trustworthiness metrics beyond accuracy. These limitations highlight the need for a paradigm shift toward data-centric approaches that directly operate in the input space and allow flexible control over multiple trust objectives. Positioned at the intersection of trustworthy AI and data valuation, VTruST addresses this gap by providing a controllable framework that enables principled trade-offs among fairness, robustness, and accuracy through targeted data subset selection in an online training setting.

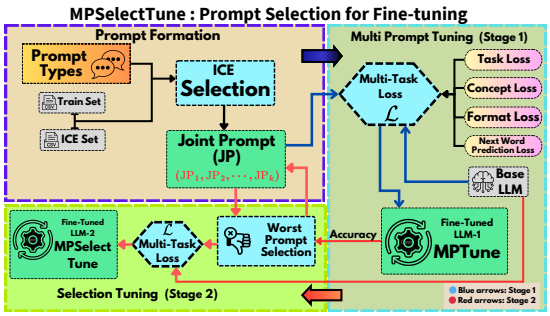
# Chapter 4

# MPSelectTune: Prompt-type Selection for Fine-tuning improves Concept Unlearning in LLMs

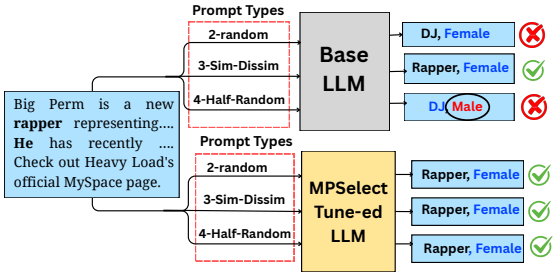
## 4.1 Introduction

LLM unlearning [24] has emerged as an important component of overall LLM safety and compliance objectives in many organizations. The LLM unlearning objective can be broadly divided into two types: (1) information unlearning (IU) [27], which erases personally identifiable information from the model, and (2) concept unlearning (CU) [7]. Concept unlearning attempts to erase the effect of a biased or harmful concept (usually in the context of a task) from the LLM, e.g. gender removal in the context of profession prediction [75] or toxicity prediction [76], removal of information about biological weapons in the context of scientific question answering [77], etc. The concept to be unlearned is specified as a dataset called the *forget set*, while *retain set* [47] provides information to be preserved (related to the main task) in the model. In this work, we focus on concept unlearning.

Concept erasure in the representation learning setup [44, 78] assumes that the concept can be represented using a linear subspace of the output representation of the features of




**(a)** Flow diagram of the proposed framework showing the main components of each stage.




**(b)** An illustrative example showing that fine-tuning using worst prompt type leads to better concept unlearning and task prediction across multiple prompt types.

**Figure 4.1** Overview of the proposed MPSelectTune framework.

the examples. However, for LLMs, zero-shot prompting techniques [79, 80], and few-shot prompting techniques involving in-context learning [81] enable various predictive tasks. In this *prompt-based predictive model* setup, the representation unlearning techniques are not directly applicable for two reasons: (1) the predictive performance of the model depends strongly on the type of prompt used to elicit concept labels, unlike in the representation learning setup, and (2) the relationship between LLM representations and predictive performance remains unclear.

In this work, we propose to use *joint task and concept prediction prompt types*, for unlearning concepts from LLMs. Fig. 4.1 (Left) shows the flow of our method. We generate multiple joint-prediction prompts for each example by varying the number and selection method of in-context examples. Stage-1 of the proposed method, called ***Multi-Prompt tuning***, uses multiple prompt types and multi-task loss for the main task and concept task while fine-tuning the model parameters. To effectively utilize the outputs of the joint prediction, we propose a novel **format loss** which forces the LLM to follow the output format for the different generated prompt types. We observe that certain prompt types accurately predict the concept labels from the fine-tuned models despite low average accuracy over all prompt types, thus demonstrating that the LLM has not truly unlearned the concept. This problem is alleviated in stage-2 of the proposed methods, called ***Selection Tuning***, where we fine-tune using the worst concept predictor prompt type. Fine-tuning using the worst prompt type is a central hypothesis of this work, since it's effectiveness towards reduction in accu-

racy of other prompt types demonstrates that the model is indeed unlearning the concept. Fig. 4.1 (Right) illustrates the effect of selection tuning, where all prompt types predict the concept label incorrectly, and the task label correctly. Experimental comparison on 5 benchmark unlearning tasks show $2 - 15\%$ points higher task prediction accuracy by the proposed method, while consistently achieving near random performance on the concept prediction task, a reduction of up to $17\%$ points compared to recent baselines. The proposed method also shows a dramatic reduction ($74\% - 23\%$) in the spurious correlation between prediction accuracies of task and concept labels using the spuriousness-score metric.

## 4.2 LLM Concept Unlearning

### 4.2.1 Problem Definition

The primary goal of LLM concept unlearning (or concept erasure) is to eliminate a specific concept encoded in a dataset from a pre-trained large language model (LLM). Such concepts may range from social biases (e.g., gender information in profession prediction [75]) to safety-critical knowledge (e.g., harmful bio-weapon–related information in scientific QA [77]). Formally, let $\mathcal{D}_c = \{(x_c(i), y_c(i)), i = 1, ..., n_c\}$ denote the dataset corresponding to the concept that must be removed (the forget set), and $\mathcal{D}_t = \{(x_t(j), y_t(j)), j = 1, ..., n_t\}$ denote the dataset for the main predictive task that the LLM system should continue to perform (the retain set). For example, in the profession prediction setting, each $x_c$ and $x_t$ corresponds to a biography text; $y_c$ denotes the gender (to be removed), whereas $y_t$ denotes the profession (to be retained). An LLM-based prediction system relies on two main components: the LLM itself, denoted by $\Theta$, and the prompt used for prediction, denoted by $\mathcal{P}$. We therefore represent the overall prediction algorithm as: $\mathcal{A} = (\Theta, \mathcal{P})$

**Instruction:** `Determine the correct answers for both questions.`

**Exemplars:** List of Exemplars – $[x_t, x_c, y_t, y_c]$

```
 Q1:  What occurs when ...
Options:  A: ...  B: ...  C: ...  D: ...
Q2:  ...
Options:  A: ...  B: ...  C: ...  D: ...
Answer:  D, D
```

*... [Repeats]*

**Test Input:**

```
 Now, solve this ...
Q1:  ...
Options:  A: ...  B: ...  C: ...  D: ...
Q2:  ...
Options:  A: ...  B: ...  C: ...  D: ...
Model Answer:  B, D
```

**Figure 4.2** Prompt structure with instruction, exemplars, and test input.

We want the prediction performance on the main task to be as high as possible, while not utilizing the concept information. We formalize this objective using the following two steps: (1) create a joint prompt $\mathcal{P}$ for solving the main task, as well as the concept prediction task, and (2) use the prompt for prediction using the LLM. Hence our predictive algorithm can be described as:

$$\hat{y}_t, \hat{y}_c = \mathcal{A}(\mathcal{P}(x_t, x_c)|\Theta) \tag{4.1}$$

where $\hat{y}_t$ and $\hat{y}_c$ are the predicted task and concept labels, respectively. The key difference between LLM concept unlearning and representation-based concept unlearning [44] is that the prompt $\mathcal{P}$ plays a key role in predictive tasks using LLMs. Hence, the unlearning objective is a joint optimization over both the prompt $\mathcal{P}$ and the LLM parameters $\Theta$. The next section describes various joint prediction prompts used for unlearning. Section 4.2.3 describes the loss functions and unlearning schemes, and Section 4.2.4 presents our complete MPSelectTune algorithm.

### 4.2.2 Joint Prediction Prompt

Figure 4.2 describes the structure of the prompt $\mathcal{P}$, with an example from the scientific QA task [77]. The prompt has 3 major sections: instruction, exemplars, and the test input. The **instruction** section includes general instructions to the LLM, followed by choices for the output(s), followed by the output format. The **exemplars** or in-context examples section provides a list of joint examples and labels from the retain and forget datasets. A joint exemplar is constructed from the task examples, $(x_t, y_t)$ from the retain set, and the concept examples $(x_c, y_c)$ from the forget set, as $(x_t, x_c, y_t, y_c)$. Finally, the **test input** section provides instructions to the LLM for solving the test question, followed by the test examples from the task, the concept, and a model answer. Generally, the **joint exemplars** (JE) are created by randomly pairing examples from the retain set $\mathcal{D}_t$ with those from the forget set $\mathcal{D}_c$. However, some tasks (e.g. profession prediction) come with a single joint example $(x_t = x_c, y_c, y_t)$. A fixed number of joint exemplars, say $k$ (which is a hyperparameter), are selected for construction of the joint prompt $\mathcal{P}$. From the given input datasets $\mathcal{D}_c$ (forget set) and $\mathcal{D}_t$ (retain set), we construct an expanded dataset of joint exemplars, which are further divided into 3 disjoint datasets of joint exemplars: (i) $\mathcal{D}_{ICE}$ - In-context exemplars, (ii) $\mathcal{D}_{tr}$ - training dataset, and (iii) $\mathcal{D}_{ts}$ - test dataset.

For constructing the prompt corresponding to a given joint example $(x'_t, x'_c, y'_t, y'_c) \in \mathcal{D}_{tr} \cup \mathcal{D}_{ts}$, joint exemplars (JEs) are selected using one of two strategies: (1) based on the cosine similarity between the embeddings of the example $(x'_t, x'_c)$ and the in-context exemplars $(x_t, x_c) \in \mathcal{D}_{ICE}$, or (2) by sampling randomly from the pool of all JEs. We use *SentenceTransformer* [82] to compute similarity scores between test inputs and exemplars. In the similarity-based selection setting, prior work has shown that maintaining diversity among exemplars can improve prediction performance [83]. To incorporate this, we use three simple approaches for prompt building: (i) `sim_dissim`: select 50% of exemplars with the highest similarity to the test input, and 50% with the lowest similarity, (ii) `half_random`: select 50% of exemplars with the highest similarity scores, and choose the remaining 50% at random, and (iii) `random`: all exemplars are selected at random. Each of these prompt building approaches, along with a size $k$ (number of exemplars), constitutes a *prompt type*. We consider 4 prompt sizes, $k = 2, 3, 4, 5$, thus constituting a total of 12 *prompt types*. Note that the actual generated prompt also depends on the joint

example. Unlike ICUL [27], which constructs exemplars by flipping concept labels $y_c$ in exemplars, our method solely uses exemplar selection strategies for prompt construction. Algorithm 3 describes generation of expanded dataset and prompt construction in detail.

**Algorithm 3:** Prompt Generation

**Input:** Concept dataset $\mathcal{D}_c$, Task dataset $\mathcal{D}_t$;
Exemplar selection methods $\{\texttt{sim\_dissim}, \texttt{random}, \texttt{half\_random}\}$;
Joint exemplars per prompt $k \in \{2, 3, 4, 5\}$

**Output:** Prompt list $\mathcal{P}_{\text{list}}$, Expanded dataset $\mathcal{D}_{\text{expanded}}$

Initialize: $\mathcal{P}_{\text{list}} \leftarrow \emptyset$;
**for** *each selection method* $s \in \{\texttt{sim\_dissim}, \texttt{random}, \texttt{half\_random}\}$ **do**
  **for** *each exemplar count* $k \in \{2, 3, 4, 5\}$ **do**
    Generate prompt template $\mathcal{P}_{s,k}$ as described in Section 4.2.2;
    Add to prompt list: $\mathcal{P}_{\text{list}} \leftarrow \mathcal{P}_{\text{list}} \cup \{\mathcal{P}_{s,k}\}$;
Create expanded dataset: $\mathcal{D}_{\text{expanded}} \leftarrow \{(x_t, y_t, x_c, y_c, \mathcal{P}_i) : (x_t, y_t, x_c, y_c) \in \mathcal{D}_t \otimes \mathcal{D}_c, \mathcal{P}_i \in \mathcal{P}_{\text{list}}\}$;
**return** $\mathcal{P}_{list}$, $\mathcal{D}_{expanded}$

### 4.2.3 Loss functions for Concept Unlearning

Given a joint training dataset $\mathcal{D}_{tr}$, we generate a list of prompts $Plist$ for each training example using the above-defined prompt types: $Plist = [\mathcal{P}_1, ..., \mathcal{P}_m]$, where $m = 12 \times |\mathcal{D}_{tr}|$. The next key steps towards developing an LLM concept unlearning algorithm is to define various loss functions corresponding to each of the prompts, and then optimize the total loss w.r.t. the LLM parameter $\Theta$. In most LLM concept unlearning tasks, there are 3 objectives: (1) minimize the loss over the primary prediction task $L_T(\Theta|Plist)$, called **task loss**, (2) minimize the **next-word-prediction (NWP) loss** $L_G(\Theta|\mathcal{D}_{tr})$ for retaining the ability of the Causal LLM for general purpose tasks, e.g. language understanding tasks [84], and (3) randomize the concept label prediction using the **concept loss** $L_C(\Theta|Plist)$. The task loss and the concept loss depend on the prompt $\mathcal{P}$, while the NWP is a standard loss over

the text in joint examples of $\mathcal{D}_{tr}$.The **task loss** is defined as:

$$L_T(\Theta|Plist) = \frac{1}{m} \sum_{\mathcal{P}_i \in Plist} l(y_t, \mathcal{A}_t(\mathcal{P}_i|\Theta)) \tag{4.2}$$

where $l$ is a standard classification loss (e.g., cross-entropy) applied to the predicted task label $\mathcal{A}_t(\mathcal{P}_i|\Theta)$, from LLM $\Theta$ and prompt $\mathcal{P}_i$.

The **concept loss** function is designed to randomize concept predictions, effectively preventing the model from learning spurious concept-task correlations. It is defined as:

$$L_C(\Theta|Plist) = 1 - \sigma(L'_C(\Theta|Plist)) \tag{4.3}$$

where $\sigma(a) = \frac{1}{1+e^{-a}}$ is the sigmoid function, and $L'_C(\Theta|\mathcal{P}, \mathcal{D}_c)$ is defined analogously to the task loss as: $L'_C(\Theta|Plist) = \frac{1}{m} \sum_{\mathcal{P}_i \in Plist} l(y_c, \mathcal{A}_c(\mathcal{P}_i|\Theta))$, $\mathcal{A}_c(\mathcal{P}_i|\Theta)$ being the concept label predictor from LLM $\Theta$ and prompt $\mathcal{P}_i$. Here, the key idea is to maximize a squashed version of the concept target prediction loss $L'_C$, thus effectively leading to randomization of the concept prediction output.

**Format loss:** While fine-tuning, we observed that the output tokens generated by the LLM do not always follow the intended format, leading to unstable behavior while calculating the task and concept loss. To fix this issue, we define the **format loss** $L_F(\Theta|Plist)$, which penalizes the format violation. Let $j \in \{1, ..., N\}$ represent a position in the token generation window, with $N$ being the maximum window length. Also, let $k \in \{1, ..., V\}$ denote the indices over the vocabulary of size $V$. We define the mask function $M_{j,k}$, as $M_{j,k} = 1$ if the $k^{th}$ token at position $j$ follows the correct format, $0$ otherwise. This is implemented using a regular expression and identifying the allowed tokens for each label. Let $P_{j,k}(i)$, the LLM generated probability of token $k$ at position $j$ defined as: $P_{j,k}(i) = \frac{\exp(\text{logits}(\mathcal{P}_i|\Theta)_{j,k})}{\sum_{l=1}^{V} \exp(\text{logits}(\mathcal{P}_i|\Theta)_{j,l})}$, where $\text{logits}(\mathcal{P}_i|\Theta)_{j,k}$ are the raw outputs generated by the LLM with prompt $\mathcal{P}_i$ for position $j$ and token $k$. The probability of a valid token at position $j$ can be computed as $VP(j,i) = \sum_{k=1}^{V} M_{j,k} \cdot P_{j,k}(i)$. We define the format loss $l$ for a given input joint prompt $\mathcal{P}_i$ as:

$$l(\mathcal{P}_i|\Theta) = -\frac{1}{N} \sum_{j=1}^{N} \log\left(VP(j,i) + \epsilon\right) \tag{4.4}$$

Finally, the total format loss can be calculated as:

$$L_F(\Theta|Plist) = \frac{1}{m} \sum_{\mathcal{P}_i \in Plist} l(\mathcal{P}_i|\Theta) \tag{4.5}$$

Combining all the losses for a multi-task learning setup, we derive the total loss function for our first proposed method, **MPTune** (Multi-prompt fine-tuning), for a prompt $\mathcal{P}$ as:

$$\mathcal{L}(\Theta|\mathcal{D}_{tr}, Plist) = \eta_T L_T(\Theta|Plist) + \eta_C L_C(\Theta|Plist) + \eta_G L_G(\Theta|\mathcal{D}_{tr}) + \eta_F L_F(\Theta|Plist) \tag{4.6}$$

where $\eta_T, \eta_C, \eta_G, \eta_F$ are weights for the different tasks in the multi-task objective. The objective for MPTune is defined as:

$$\Theta^{\text{MPTune}} = \arg\min_{\Theta} \mathcal{L}(\Theta|\mathcal{D}_{tr}, Plist) \tag{4.7}$$

This objective can be efficiently optimized using LoRa fine-tuning [85] for state-of-the-art LLMs.

### 4.2.4 Prompt-type selection for LLM Concept Unlearning

The objective in equation 4.7 is to provide equal weightage to all the 12 prompt types. However, we observe (from results in section 4.3.3) that some prompt types perform poorly in terms of concept unlearning, compared to other prompts. In other words, the accuracy of concept prediction using certain prompt types can go up to $\sim 71\%$, even though the average accuracy is less than $60\%$, for an unlearned MPTune model. In this section, we propose **MPSelectTune**, which addresses this key limitation of **MPTune**. More generally, the adversarial formulation of concept unlearning [44] postulates that the worst concept predictor using the unlearned representation (one having the highest accuracy) should perform poorly. We extend this notion to prompt-types in the case of LLM concept unlearning as: *the concept prediction accuracy of the worst prompt-type (with highest accuracy) should be minimized.*

This objective, called **MPSelectTune**, can be formalized as:

$$\Theta^{\text{MPSelectTune}} = \arg\min_{\Theta} \mathcal{L}(\Theta|\mathcal{D}_{tr}, Plist_{sel}) \tag{4.8}$$

**Algorithm 4:** MPSelectTune: Prompt-type Selection and Fine-tuning

**Input:** Joint
training dataset $\mathcal{D}_{tr}$; Pre-trained LLM parameters $\Theta_0$; Prompt-type set $\mathcal{P}_{\text{list}}$

**Output:** Unlearned model parameters $\Theta^{\text{MPSelectTune}}$

**Prompt Generation:** Generate
a prompt list $\mathcal{P}_{\text{all}}$ by applying each prompt-type $\pi_i \in \mathcal{P}_{\text{list}}$ to every example in $\mathcal{D}_{tr}$.

**Initial Fine-tuning:** Obtain intermediate parameters $\Theta^{\text{MPTune}}$ using Eq. 4.7.

**Prompt-type Evaluation: foreach** $\pi_i \in \mathcal{P}_{list}$ **do**
    Compute concept accuracy $\text{Acc}_c(\pi_i \mid \mathcal{D}_{tr}, \Theta^{\text{MPTune}})$;

**Prompt-type Selection:** Select the worst-performing prompt-type:

$$\pi^* = \arg\max_{\pi_i \in \mathcal{P}_{\text{list}}} \text{Acc}_c(\pi_i)$$

**Selective Prompt Construction:** Construct $\mathcal{P}_{\text{sel}}$ using $\pi^*$ for all examples in $\mathcal{D}_{tr}$.

**Final Fine-tuning:** Optimize the model parameters:

$$\Theta^{\text{MPSelectTune}} = \arg\min_{\Theta} \mathcal{L}(\Theta \mid \mathcal{D}_{tr}, \mathcal{P}_{\text{sel}})$$

**return** $\Theta^{MPSelectTune}$

where $Plist_{sel}$ is the list of prompts generated from the training dataset $\mathcal{D}_{tr}$ using the highest-accuracy prompt type $\pi^*$ This leads us to a two-stage scheme where stage 1 computes $\Theta^{\text{MPTune}}$ using the multi-task and multi-prompt-type loss function, and stage 2 uses the worst prompt-type from stage 1, to further fine-tune the model parameters to compute $\Theta^{\text{MPSelectTune}}$. The complete algorithm is described in Algorithm. 4.

## 4.3 Experimental Results

In this section, we describe the experimental results comparing the proposed method MPSelectTune with several state-of-the-art baselines. Our primary **research question** is:

*Can fine-tuning with worst prompt type effectively unlearn a concept from LLM?* Section 4.3.1 describes the experimental setup, while section 4.3.2 compares the performances of the proposed methods with baselines and tries to answer the primary research question. Sections 4.3.3 and 4.3.4 further analyses the prompt type-specific performance and components of the multi-task loss. Finally, Section 4.3.5 provides the generalization of the proposed method on unseen prompt types.

### 4.3.1 Experimental Setup

**Datasets:** We use 5 task-concept-pair Dataset to evaluate our method on LLaMA2-7B, LLaMA3.1-8B, and Mistral-7B-Instruct-v0.3. For **Bios** [75], **RT-Gender** [86], and **ToxicBias** [76], the main tasks are *profession, sentiment, and toxicity* prediction, while the concept task is *gender* prediction. **Adult Census** [66] predicts income level (exceeds $50K) as main task and *race* as concept. **SciQ-WMDPBio** combines scientific question-answering [87] as main task with bio-weapons QA [77] as concept task. WMDPBio was used in [7] for concept unlearning evaluation. This combination provides the hardest unlearning scenario as SciQ and WMDPBio tasks are similar.

**Metrics:** We assess our method and baselines along four dimensions. **(1) main task accuracy** (Task-Acc) and **(2) concept accuracy** (Concept-Acc) form the primary evaluation components with high main task accuracy and near-random concept accuracy being the most desirable. **3. MMLU Accuracy** (MMLU-Acc): We also evaluate the unlearned models' performance on the standard MMLU benchmark dataset [84], in order to ensure that the unlearning process does not generic model performance (unrelated to the main task).

**4. Spuriousness Score (SP-Score)**: This metric was proposed in [88] for determining whether the spurious correlations between the main task labels and the concept labels are utilized by a given classifier. In the binary classification setting, the *minor group* is defined as the pair of main task and concept task labels that are not expected to be spuriously correlated. The spuriousness score was defined as: $|1 - \frac{Acc_f}{Acc_c}|$ where $Acc_f$ is the accuracy of the given classifier $f$ on the minor group, and $Acc_c$ is the accuracy of a "clean" classifier

(one without spurious correlation), on the minor group. A higher spuriousness score denotes a relatively lower accuracy of the given classifier on minor group, thus signifying a higher reliance of the classifier $f$ on spuriously related concept labels.

We generalize the spuriousness score metric to the setting where the main task is multi-class classification. For the construction of minority sets, each main task label is annotated to have a corresponding spurious concept label. For the profession prediction task, `(Nurse, Female)` and `(Doctor, Male)` can be spuriously correlated pairs. The minor set $S_{\text{minor}}$ is constructed as all non-spuriously correlated pairs of labels, e.g., `(Nurse, Male)`, `(Doctor, Female)`. We define the *SP-Score* as:

$$\text{SP-Score}(f) = \max_{i \in \{\text{M,F}\}} \left| 1 - \frac{\text{Acc}_f}{\text{Acc}_{c_i}} \right| \tag{4.9}$$

where $\text{Acc}_f$ is the task accuracy of the given model $f$ on $S_{\text{minor}}$, and $\text{Acc}_{c_i}$ is the task accuracy of the clean model $c_i$. In our (in-context learning) setting, the different models, $f, c_M, c_F$ are distinguished by the in-context examples used in prompts. The model $f$ uses the entire set of selected in-context examples as described in section 4.2.2. The "clean" models $c_M$ and $c_F$, only use in-context examples with concept labels `Male` and `Female`, respectively. Other selection criteria remain unchanged. This procedure is analogous to [88], except that we use clean classifiers constructed from both male and female classes, whereas they only use one of them. We find that due to lower influence of the dataset on in-context learning (compared to model training), the values of *SP-Score* are lower in our setting. Hence, taking the maximum over $M$ or $F$ gives us a more robust score, which considers the "cleaner" of the two base classifiers.

**Baselines:** We benchmark our approach against unlearning algorithms using both pre-LLM representation unlearning models and LLM-based baselines with LLaMA2-7B, LLaMA3.1-8B, and Mistral-7B-Instruct-v0.3. **Pre-LLM baselines** include pre-trained *BERT-base* embeddings [22], *KRAM* [46], *RLACE* [44], and *KCE* [45]. **LLM-based baselines** include the base models (*Base*), the fine-tuned model using 12 sets of prompt types across all custom datasets with all retained labels (*FT*), and the augmented fine-tuned model with flipped concept labels (*Aug*). Fine-tuning is performed using Low-Rank Adaptation (*LoRA*) [89] with rank $= 8$ and $\alpha = 64$. Additionally, we benchmark against recent state-of-the-art methods: *ICUL* [27] and *SKU* [25], where SKU is a gradient-based

method for machine unlearning. For the **SciQ-WMDP-Bio** dataset, we also compare against the SOTA *ECK* baseline [7].

**Proposed Method:** Our proposed approach consists of two stages: Both **MPTune** (Stage 1) and **MPSelectTune** (Stage 2) fine-tune the base model with the multi-task loss from Section 4.2.3: Stage 1 uses all prompt types, while Stage 2 focuses on the worst prompt type for robust concept unlearning.

**Table 4.1** Comparison of unlearning performance on BIOS and RT-Gender datasets with LLM-based Baselines. The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics).

| Method | Task Acc | Concept Acc | MMLU Acc | SP Score | Task Acc | Concept Acc | MMLU Acc | SP Score |
|---|---|---|---|---|---|---|---|---|
| | **Bios Dataset** | | | | **RT-Gender Dataset** | | | |
| **Model: Llama-2-7B** | | | | | | | | |
| Base | 89.50 | 93.40 | 43.9 | 0.132 | 58.54 | 71.30 | 43.9 | 0.146 |
| FT | 99.82 | 99.96 | 42.1 | *0.019* | 70.08 | 86.42 | 40.2 | *0.043* |
| Aug | 95.04 | 92.81 | 37.6 | 0.065 | 64.17 | 82.50 | 37.6 | 0.108 |
| ICUL | *84.36* | 83.64 | 42.1 | 0.185 | *67.43* | 73.25 | 40.2 | 0.118 |
| SKU | 72.75 | *65.55* | 34.9 | 0.302 | 65.36 | *59.45* | 37.4 | 0.121 |
| **MPTune** | **99.82**(+15.5%) | **61.57**(−4.0%) | 42.8 | **0.012** | **70.00**(+2.6%) | **53.83**(−5.6%) | 42.6 | 0.021 |
| **MPSelectTune** | **99.79**(+15.4%) | **55.6**(−10.0%) | 42.9 | **0.011** | **70.08**(+2.7%) | **51.50**(−8.0%) | 43.1 | **0.011** |
| **Model: Llama-3.1-8B** | | | | | | | | |
| Base | 90.14 | 96.33 | 65.0 | 0.100 | 63.39 | 75.36 | 65.0 | 0.173 |
| FT | 99.43 | 98.7 | 63.1 | *0.030* | 71.12 | 86.87 | 59.6 | *0.056* |
| Aug | 97.46 | 88.76 | 58.9 | 0.052 | 67.31 | 77.35 | 59.7 | 0.123 |
| ICUL | *87.46* | *73.86* | 63.1 | 0.149 | *64.22* | *66.93* | 59.6 | 0.144 |
| SKU | 78.32 | 74.86 | 31.9 | 0.225 | 73.58 | 67.33 | 61.9 | 0.105 |
| **MPTune** | **99.36**(+11.9%) | **59.36**(−14.5%) | 64.2 | **0.017** | **70.96**(+6.7%) | **54.33**(−12.6%) | 64.4 | **0.029** |
| **MPSelectTune** | **99.25**(+11.8%) | **56.61**(−17.3%) | 64.3 | **0.019** | **71.03**(+6.8%) | **49.81**(−17.1%) | 64.2 | **0.032** |
| **Model: Mistral-7B-Instruct-v0.3** | | | | | | | | |
| Base | 84.4 | 91.9 | 56.1 | 0.129 | 59.2 | 72.8 | 56.1 | 0.146 |
| FT | 96.6 | 94.7 | 53.7 | *0.025* | 68.9 | 85.3 | 53.8 | *0.043* |
| Aug | 93.8 | 90.2 | 52.6 | 0.067 | 65.4 | 80.7 | 54.0 | 0.108 |
| ICUL | *90.3* | *67.4* | 53.7 | 0.101 | *66.8* | *71.6* | 53.8 | 0.118 |
| SKU | 75.6 | 68.9 | 52.3 | 0.156 | 64.7 | 65.2 | 52.3 | 0.121 |
| **MPTune** | **92.5**(+2.2%) | **63.8**(−3.6%) | 54.0 | **0.023** | **69.1**(+2.3%) | **55.7**(−15.9%) | 53.8 | **0.023** |
| **MPSelectTune** | **92.1**(+1.8%) | **61.1**(−6.3%) | 53.8 | **0.021** | **69.3**(+2.5%) | **52.4**(−19.2%) | 53.7 | **0.021** |

**Table 4.2** Comparison of unlearning performance on Toxic Bias and Adult Census datasets with LLM-based Baselines. The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics).

| Method | Task Acc | Concept Acc | MMLU Acc | SP Score | Task Acc | Concept Acc | MMLU Acc | SP Score |
|---|---|---|---|---|---|---|---|---|
| | **Toxic Bias Dataset** | | | | **Adult Census Dataset** | | | |
| **Model: Llama-2-7B** | | | | | | | | |
| Base | 75.41 | 82.25 | 43.9 | 0.116 | 62.2 | 57.6 | 43.9 | 0.260 |
| FT | 89.92 | 95.67 | 41.1 | *0.050* | 75.6 | 71.2 | 36.8 | 0.121 |
| Aug | 81.46 | 86.33 | 39.4 | 0.135 | 68.4 | 67.7 | 36.9 | 0.197 |
| ICUL | *86.50* | *66.96* | 41.1 | 0.056 | *70.9* | *61.4* | 36.8 | 0.151 |
| SKU | 80.46 | 68.33 | 38.6 | 0.114 | 69.7 | 62.6 | 37.0 | 0.170 |
| **MPTune** | **89.63**(+3.1%) | **60.17**(−6.8%) | 41.9 | **0.028** | **74.9**(+4.0%) | **58.4**(−3.0%) | 36.2 | 0.079 |
| **MPSelectTune** | **89.75**(+3.3%) | **53.13**(−13.8%) | 42.0 | **0.026** | **74.7**(+3.8%) | **57.6**(−3.8%) | 35.9 | **0.068** |
| **Model: Llama-3.1-8B** | | | | | | | | |
| Base | 77.66 | 83.41 | 65.0 | 0.166 | 68.6 | 59.4 | 65.0 | 0.261 |
| FT | 90.12 | 94.33 | 61.7 | 0.030 | 79.3 | 73.7 | 61.8 | *0.116* |
| Aug | 84.36 | 75.17 | 58.3 | 0.119 | 75.9 | 64.3 | 60.3 | 0.185 |
| ICUL | *81.35* | *65.97* | 61.7 | 0.134 | *72.4* | *59.8* | 61.8 | 0.214 |
| SKU | 80.63 | 69.42 | 60.3 | 0.156 | 70.6 | 61.7 | 60.3 | 0.187 |
| **MPTune** | **90.06**(+8.7%) | **64.12**(−1.9%) | 62.1 | 0.023 | **78.0**(+5.6%) | **59.2**(−0.6%) | 61.9 | **0.074** |
| **MPSelectTune** | **89.93**(+8.6%) | **58.34**(−7.6%) | 62.8 | **0.016** | **77.7**(+5.3%) | **56.9**(−2.9%) | 62.0 | **0.079** |
| **Model: Mistral-7B-Instruct-v0.3** | | | | | | | | |
| Base | 76.3 | 83.1 | 56.1 | 0.129 | 61.4 | 58.2 | 56.1 | 0.260 |
| FT | 88.7 | 94.8 | 53.4 | *0.025* | 74.8 | 70.6 | 54.0 | 0.121 |
| Aug | 82.1 | 84.9 | 52.8 | 0.067 | 67.9 | 65.8 | 53.1 | 0.197 |
| ICUL | *85.2* | *68.3* | 53.4 | 0.101 | *70.1* | *60.8* | 54.0 | 0.151 |
| SKU | 79.8 | 70.1 | 50.3 | 0.156 | 68.9 | 61.9 | 52.0 | 0.170 |
| **MPTune** | **88.4**(+3.2%) | **62.5**(−5.8%) | 53.2 | **0.023** | **73.6**(+3.5%) | **59.2**(−1.6%) | 52.8 | **0.079** |
| **MPSelectTune** | **88.6**(+3.4%) | **56.8**(−11.5%) | 53.4 | **0.021** | **73.4**(+3.3%) | **57.8**(−3.0%) | 53.1 | **0.068** |

### 4.3.2 Comparison of Unlearning Performance

Table 4.1 and Table 4.2 reports results comparing MPTune and MPSelectTune with LLM-based baselines, for datasets Bios, RT-Gender, ToxicBias, and Adult Census. Note that all the metrics reported are averaged over all prompt types. Across all datasets, MPTune and MPSelectTune consistently achieve main task accuracy comparable to the FT model while reducing concept task accuracy to near-random. MPSelectTune is especially effective at unlearning in terms of average concept accuracy, despite being fine-tuned for the worst-case prompt type. This validates the central hypothesis of this work: *fine-tuning using the worst-case prompt type removes the concept from the LLM more effectively*. Both methods

maintain MMLU accuracy close to their respective base models, within 2% for LLaMA-2 and 3% for LLaMA-3.1 and Mistral. In terms of SP-score, our methods outperform all baselines with a significant margin of 23–74%. This further validates our hypothesis that fine-tuning with worst-case prompt type removes spurious correlations between the concept and the main task, thus enabling the LLM to predict without using concept.

Table 4.3 presents a comparison of our proposed methods with pre-LLM baselines across three datasets. Notably, our methods not only approach but often surpass the performance of traditional representation unlearning approaches, demonstrating superior concept unlearning while maintaining strong task accuracy.

**Table 4.3** Performance comparison with Pre-LLM baselines (representation unlearning). The values in brackets show percentage point improvement (+ for main task and − for concept) over the closest baseline (in italics).

| **Method** | **Bios Dataset** | | **RT-Gender Dataset** | | **ToxicBias Dataset** | |
|---|---|---|---|---|---|---|
| | **Task-Acc** | **Concept-Acc** | **Task-Acc** | **Concept-Acc** | **Task-Acc** | **Concept-Acc** |
| Bert-base | 79.47 | 89.06 | 67.29 | 73.68 | 69.21 | 72.58 |
| KRAM [46] | *76.82* | *62.86* | 55.17 | *61.13* | 65.33 | *64.89* |
| RLACE [44] | 61.2 | 65.92 | 62.2 | 67.8 | *68.00* | 65.33 |
| KCE [45] | 56.08 | 63.94 | *66.30* | 68.20 | 67.33 | 66.72 |
| **Model: Llama-3.1** | | | | | | |
| **MPTune** (Proposed) | 99.36(+22.5%) | 59.36(−3.5%) | 70.96(+4.7%) | 54.33(−6.8%) | 90.06(+22.1%) | 64.12(−0.8%) |
| **MPSelectTune** (Proposed) | 99.25(+22.4%) | 56.61(−6.3%) | 71.03(+4.7%) | 49.81(−11.3%) | 89.93(+21.9%) | 58.34(−6.6%) |

For the SciQ-WMDP-Bio dataset, the results in Table 4.3 show that our methods achieve a substantial reduction in concept accuracy while preserving task accuracy and MMLU performance, validating the robustness of our approach across different model architectures.

**Table 4.4** Unlearning on SciQ-WMDP-Bio Dataset using Llama-2, Llama-3.1, and Mistral-7B

| **Method** | **Llama-2** | | | **Llama-3.1** | | | **Mistral-7B** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Task-Acc** | **Concept-Acc** | **MMLU-Acc** | **Task-Acc** | **Concept-Acc** | **MMLU-Acc** | **Task-Acc** | **Concept-Acc** | **MMLU-Acc** |
| Base | 23.1 | 19.7 | 43.9 | 68.4 | 61.3 | 65.0 | 38.5 | 35.3 | 56.1 |
| FT | 25.4 | 26.1 | 24.6 | 76.5 | 68.7 | 63.8 | 42.6 | 41.2 | 52.7 |
| Aug | 21.7 | 19.6 | 26.7 | 74.6 | 42.4 | 56.6 | 41.0 | 40.0 | 51.5 |
| ECK | – | – | – | – | 32.2 | 61.6 | – | – | – |
| **MPTune** | 25.4 | 25.4 | 24.0 | **75.6** | **31.8** (–0.4%) | 64.1 | **41.5** | **33.0** | 51.7 |
| **MPSelectTune** | 24.8 | 25.1 | 24.3 | **75.4** | **29.9** (–2.3%) | 64.3 | **40.7** | **30.3** | 52.1 |

In summary, MPTune and MPSelectTune effectively unlearn concept information

while retaining task-specific and general language capabilities better than all considered baselines. Next, we observe the performance of the methods in a more granular way through the lens of the different prompt types.

### 4.3.3 Analysis of Prompts

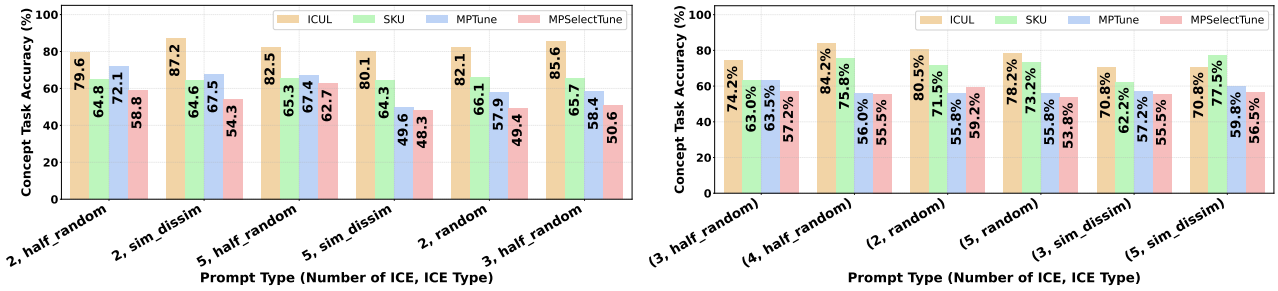



**Figure 4.3 Left**: Concept Accuracy for few prompt-types. **Right**: Generalization of the proposed methods to unseen prompt types.

Figure 4.3 (left) compares concept task accuracies for ICUL, SKU, MPTune, and MPSelectTune on 6 prompt types, 3 with highest concept accuracy using MPTune, and 3 with lowest concept accuracy using MPTune, using LLaMA-2 on the BIOS dataset. ICUL has the highest concept accuracy among all methods and for all prompt types, with its worst at 85.6% on '3, half_random'. SKU performs decently well (64.3–66.1%) on concept unlearning, at the cost of low task accuracy. MPTune achieves concept accuracies similar to SKU but with a high standard deviation of 5.51 across different prompt types. The best-performing prompt turns out to be '`5, sim_dissim`' (with 49.6% concept accuracy) and the worst-performing prompt type turns out to be '`2, half_random`' (with 72.1% concept accuracy). MPSelectTune performs uniformly better than baseline methods, with a noticeable drop in the peak concept task accuracy 62.7% across prompt types with a reduced standard deviation of 4.35.

### 4.3.4 Ablation study of loss functions

Table 4.5 reports an ablation study to assess the impact of each component in MPSelectTune's loss function. The total loss ($\mathcal{L}$) includes task prediction loss, concept prediction

**Table 4.5** Ablation of loss function components in MPSelectTune on Bios Dataset with Llama-2

| Config | Task | Concept | MMLU | SP |
|---|---|---|---|---|
| Total Loss | **99.79** | **55.6** | 42.9 | **0.011** |
| -Format L | 96.14 | 71.82 | 42.8 | 0.053 |
| -Task L | 89.46 | 63.44 | 43.0 | 0.110 |
| -Concept L | 99.11 | 98.79 | 42.2 | 0.028 |

loss, format loss, and the next-word prediction loss. As expected, removing the task loss (-Task L) reduces task accuracy by 10.33%, while ablating the concept loss (-Concept L) increases the concept accuracy by 42.19%. The relatively lower impact of task loss is due to the next word prediction loss. Removing the format loss (-Format L) raises concept accuracy by 15.22%. However, we observed that the actual prediction of the second output token is often something different from the expected tokens (e.g. Male/Female). The increase in accuracy is due to higher output probabilities of the correct token among the allowed concept tokens. In summary, all the loss components are important for generation of correct outputs.

### 4.3.5 Generalization to Unseen Prompt Types

To assess generalization, we trained our models using half of the available prompt types and evaluated using the remaining, unseen prompt types using LLaMA-2 on BIOS dataset. Figure 4.3 (right) shows the concept accuracies for the remaining prompt types. ICUL consistently shows the highest concept accuracy on unseen prompts (76.5% average), indicating limited generalization. SKU's concept accuracy is the second highest on unseen prompts (70.5% average), compared to its overall average when trained and tested on all prompts (65.6%), showing a lack of generalization. In contrast, both the proposed methods achieve low concept accuracies, demonstrating better generalization. MPTune and MPSelectTune obtain 58.0% and 56.3% average concept accuracy on unseen prompts, respectively, which are similar to than their averages when trained and tested on all prompts (61.6% and 55.5%). The slight decrease in MPTune's concept accuracy is due to the fact that the chosen prompt-types had higher than average concept accuracy. These

results show that both the proposed methods generalize effectively to new prompt types.

## 4.4 Conclusions

In this work, we explore the design of an adversarial prompt-based fine-tuning for unlearning concepts from an LLM. We propose a two stage approach called *MPSelectTune*, that uses a multi-task loss function to fine-tune the LLMs for unlearning using the worst prompt. Our experiments demonstrate that the proposed method is successful in outperforming several recent state-of-the-art baselines, thus highlighting their efficacy in the area of concept unlearning or concept erasure.

# Chapter 5

# Conclusion and Future Work

This thesis addresses two fundamental challenges in trustworthy machine learning: (i) designing data-centric mechanisms to control trade-offs among multiple trustworthiness objectives such as accuracy, fairness, and robustness, and (ii) developing robust concept unlearning strategies for large language models (LLMs) that generalize across diverse prompting conditions. Through two complementary contributions, this thesis demonstrates that both data selection and prompt selection are powerful levers for building trustworthy and controllable AI systems. In this chapter, we summarize the key contributions of the thesis and discuss promising future research directions motivated by this work.

## 5.1 Summary of Contributions

The major contributions of this thesis can be broadly categorized into two parts: (1) controllable data subset selection for trustworthy learning, and (2) prompt-type selection for robust concept unlearning in LLMs.

### 5.1.1 Controllable Data Subset Selection for Trustworthy AI (VTruST)

To address the challenge of balancing competing trustworthiness objectives, this thesis proposes **VTruST**, a data-centric framework for controllable training data subset selection.

- We introduce a unified value-function-based formulation that quantifies the contribution of individual training samples toward multiple trustworthiness metrics, including accuracy, fairness, and robustness. These value functions are designed to be additive, enabling principled combination through user-defined trade-off parameters.
- We formulate training data subset selection as an *online sparse approximation* problem and propose a novel **online Orthogonal Matching Pursuit (OMP)**-based algorithm. Unlike prior data valuation methods that require full dataset passes or expensive retraining, VTruST performs efficient online replacement of training samples while maintaining a fixed subset size.
- The proposed framework enables explicit control over trade-offs between trustworthiness objectives, such as accuracy–fairness and fairness–robustness, by tuning a single hyperparameter. This provides practitioners with a transparent and interpretable mechanism to explore Pareto-optimal solutions.
- Extensive experiments on social, image, and scientific datasets demonstrate that models trained on VTruST-selected subsets consistently outperform state-of-the-art baselines. while using significantly fewer training samples.
- Beyond performance gains, VTruST also provides *data-centric explanations* by identifying influential training samples for different trustworthiness objectives, thereby improving interpretability and auditability of the training process.

### 5.1.2 Prompt-Type Selection for Robust Concept Unlearning in LLMs (MPSelectTune)

While VTruST focuses on data-centric trustworthiness, the second contribution of this thesis addresses the emerging problem of **concept unlearning in large language models**.

- We identify a key limitation of existing LLM unlearning approaches: their vulnerability to prompt variation. We show that a model may appear to have unlearned a concept on average, while still retaining the concept under certain prompt types.
- To address this issue, we propose a novel two-stage framework consisting of **Multi-Prompt Tuning (MPTune)** and **Prompt-Type Selection Tuning (MPSelectTune)**. In the first stage, the model is fine-tuned using multiple prompt types and a multi-task loss that jointly optimizes task accuracy and concept randomization.
- In the second stage, we introduce an adversarial prompt-type selection strategy that identifies the *worst-case prompt type*, the one yielding the highest concept prediction accuracy, and further fine-tunes the model using only this prompt type. This ensures robust concept unlearning across all prompts.
- Experimental results on multiple benchmarks, model architectures, and datasets demonstrate that MPSelectTune consistently improves main task accuracy (by up to 15%) while reducing worst-case concept accuracy by up to 17% compared to recent baselines.
- The proposed method also significantly reduces spurious correlations between task and concept predictions, as measured by the spuriousness score, while preserving general language understanding ability as verified through MMLU evaluations.

## 5.2 Future Work

This thesis opens several promising avenues for future research, spanning data-centric learning, LLM alignment, and trustworthy AI.

- **Extending VTruST to LLM Fine-tuning:** While VTruST is evaluated on classical machine learning and deep learning models, a natural extension is to apply value-function-based subset selection to large-scale LLM fine-tuning. This could significantly reduce computational costs while enabling controllable trade-offs between safety, fairness, and task performance during instruction tuning.

- **Automatic Prompt-Type Discovery and Selection:** MPSelectTune currently operates on a predefined set of prompt types. An important direction is to develop automatic prompt generation and clustering methods that discover adversarial prompt families, enabling fully automated and stronger unlearning guarantees.

- **Multi-Class and Multi-Concept Unlearning:** The current formulation of MPSelectTune focuses on binary or single-concept unlearning. Extending the framework to handle multi-class and multi-concept settings—where multiple sensitive attributes or harmful concepts must be unlearned simultaneously—remains an open challenge.

- **Unified Data and Prompt Selection Frameworks:** A particularly exciting direction is to combine the ideas from VTruST and MPSelectTune into a unified framework that jointly selects training data subsets and prompt types. Such a framework could provide end-to-end controllability over trustworthiness in both data-centric and prompt-centric learning paradigms.

- **Theoretical Guarantees and Certification:** Finally, future research can focus on developing theoretical guarantees for value-function-based data selection and prompt-type-based unlearning, including robustness bounds and certification of concept removal.

In conclusion, this thesis demonstrates that principled selection of data in classical learning settings and of prompts in LLM-based systems plays a central role in achieving trustworthy, interpretable, and robust AI. We hope that the ideas and frameworks introduced in this work will inspire further research at the intersection of data-centric AI, LLM safety, and trustworthy machine learning.

# References


[1] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[4] M. Feldman, S. A. Friedler, *et al.*, "Certifying and removing disparate impact," in *KDD*, 2015.

[5] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *ICLR*, 2019.

[6] V. Patil, P. Hase, and M. Bansal, "Can sensitive information be deleted from llms?," in *ICLR*, 2024.

[7] R. Gandikota *et al.*, "Erasing conceptual knowledge from language models," *arXiv preprint arXiv:2410.02760*, 2024.

[8] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Computing Surveys*, vol. 55, no. 2, 2022.

[9] B. Li, P. Qi, B. Liu, *et al.*, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, 2023.

[10] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric artificial intelligence: A survey," *arXiv preprint arXiv:2303.10158*, 2023.

[11] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[12] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*, 2017.

[13] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data," in *ICML*, 2019.

[14] B. Mirzasoleiman *et al.*, "Coresets for data-efficient training," in *ICML*, 2020.

[15] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, A. De, and R. Iyer, "Grad-match: Gradient matching based data subset selection for efficient deep model training," in *International Conference on Machine Learning*, pp. 5464–5474, PMLR, 2021.

[16] H. Wang, C. Xiao, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang, "Augmax: Adversarial composition of random augmentations for robust training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 237–250, 2021.

[17] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," *NeurIPS*, 2021.

[18] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," in *International Conference on Learning Representations*, 2021.

[19] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *International Conference on Machine Learning*, pp. 8346–8356, PMLR, 2020.

[20] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[21] A. J. Weinstein and M. B. Wakin, "Online search orthogonal matching pursuit," in *IEEE Statistical Signal Processing Workshop*, pp. 584–587, 2012.

[22] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers," in *NAACL*, 2019.

[23] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[24] Y. Yao, X. Xu, and Y. Liu, "Large language model unlearning," *arXiv preprint arXiv:2310.10683*, 2023.

[25] Z. Liu, G. Dou, Z. Tan, Y. Tian, and M. Jiang, "Towards safer large language models through machine unlearning," *arXiv preprint arXiv:2402.10058*, 2024.

[26] J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo, "Knowledge unlearning for mitigating privacy risks in language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

[27] M. Pawelczyk, S. Neel, and H. Lakkaraju, "In-context unlearning: Language models as few-shot unlearners," in *ICML*, 2024.

[28] P. Thaker *et al.*, "Position: Llm unlearning benchmarks are weak measures of progress," *arXiv preprint arXiv:2410.02879*, 2024.

[29] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, pp. 325–333, PMLR, 2013.

[30] Y. Romano, S. Bates, and E. Candès, "Achieving equalized odds by resampling sensitive attributes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[31] P. Sattigeri, S. Hoffman, and V. Chenthamarakshan, "Fairness-aware learning through regularization," in *AAAI Conference on Artificial Intelligence*, 2022.

[32] C.-H. Chuang *et al.*, "Fair data augmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[33] T. Pang, K. Xu, Y. Dong, S. Chen, and J. Zhu, "Understanding robustness-accuracy tradeoffs via fourier analysis," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 17356–17368, PMLR, 2022.

[34] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of EMNLP*, 2020.

[35] K. Ethayarajh *et al.*, "Understanding dataset difficulty with model confidence," *Transactions of the ACL*, 2022.

[36] Y. Kwon, Y. Li, J. Kim, J. Kim, and C. D. Yoo, "Dataset cartography revisited: Mapping and diagnosing datasets with training dynamics," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[37] N. Seedat *et al.*, "Identifying incongruous data points for dataset debugging," in *NeurIPS*, 2022.

[38] T. Wang *et al.*, "Efficient data valuation using reinforcement learning," in *NeurIPS*, 2022.

[39] J. Park *et al.*, "Trak: Attributing model predictions to training data," in *ICML*, 2023.

[40] J. Yoon *et al.*, "Data valuation using reinforcement learning," in *ICML*, 2020.

[41] M. Paul *et al.*, "Deep core-sets for data-efficient learning," in *ICLR*, 2021.

[42] H. Just *et al.*, "Lava: Label-aware data valuation," in *ICML*, 2023.

[43] Y. Wu *et al.*, "Davinz: Data valuation without training," in *NeurIPS*, 2022.

[44] S. Ravfogel *et al.*, "Linear adversarial concept erasure," in *ICML*, 2022.

[45] S. Ravfogel *et al.*, "Adversarial concept erasure in kernel space," in *EMNLP*, 2022.

[46] S. Basu *et al.*, "Robust concept erasure with kram," in *NeurIPS*, 2023.

[47] J. Liu *et al.*, "Rethinking machine unlearning for large language models," *arXiv preprint arXiv:2401.01234*, 2024.

[48] H. Kassem *et al.*, "Preserving utility while unlearning in language models," in *EMNLP*, 2023.

[49] Z. Ding *et al.*, "A unified framework for dememorization in language models," *arXiv preprint arXiv:2403.00123*, 2024.

[50] Y. Zhang *et al.*, "Get: Parameter editing for knowledge removal," in *ICLR*, 2024.

[51] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[52] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[53] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[54] Z. Song, L. Liu, F. Jia, Y. Luo, G. Zhang, L. Yang, L. Wang, and C. Jia, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *arXiv preprint arXiv:2401.06542*, 2024.

[55] L. Benato, E. Buhmann, M. Erdmann, P. Fackeldey, J. Glombitza, N. Hartmann, G. Kasieczka, W. Korcari, T. Kuhr, J. Steinheimer, H. Stöcker, T. Plehn, and K. Zhou, "Shared data and algorithms for deep learning in fundamental physics," *Computing and Software for Big Science*, vol. 6, May 2022.

[56] X. Li, P. Wu, and J. Su, "Accurate fairness: Improving individual fairness without trading accuracy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14312–14320, 2023.

[57] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019.

[58] Y. Hu, F. Wu, H. Zhang, and H. Zhao, "Understanding the impact of adversarial robustness on accuracy disparity," in *International Conference on Machine Learning*, pp. 13679–13709, PMLR, 2023.

[59] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[60] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[61] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," in *Advances in Neural Information Processing Systems*, 2020.

[62] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29935–29948, 2021.

[63] S. Addepalli, S. Jain, *et al.*, "Efficient and effective augmentation strategy for adversarial training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1488–1501, 2022.

[64] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," in *ICLR*, 2021.

[65] Y. Mroueh *et al.*, "Fair mixup: Fairness via interpolation," in *International Conference on Learning Representations*, 2021.

[66] R. Kohavi *et al.*, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.," in *Kdd*, vol. 96, pp. 202–207, 1996.

[67] "Medical expenditure panel survey."

[68] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[69] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[70] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[72] N. Mu and J. Gilmer, "MNIST-C: A robustness benchmark for computer vision," *CoRR*, vol. abs/1906.02337, 2019.

[73] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019.

[74] S.-J. Huang, J.-W. Zhao, and Z.-Y. Liu, "Cost-effective training of deep cnns with active model adaptation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1580–1588, 2018.

[75] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

[76] N. Sahoo, H. Gupta, and P. Bhattacharyya, "Detecting unintended social bias in toxic language datasets," in *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (A. Fokkens and V. Srikumar, eds.), (Abu Dhabi, United Arab Emirates (Hybrid)), pp. 132–143, Association for Computational Linguistics, Dec. 2022.

[77] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, G. Mukobi, *et al.*, "The wmdp benchmark: Measuring and reducing malicious use with unlearning," in *ICML*, 2024.

[78] N. Belrose *et al.*, "Leace: Perfect linear concept erasure," *Transactions of the ACL*, 2024.

[79] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[80] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[81] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, *et al.*, "A survey on in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

[82] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.

[83] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.

[84] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[85] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[86] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov, "RtGender: A corpus for studying differential responses to gender," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[87] J. Welbl, N. F. Liu, and M. Gardner, "Crowdsourcing multiple choice science questions," in *Proceedings of the 3rd Workshop on Noisy User-generated Text* (L. Derczynski, W. Xu, A. Ritter, and T. Baldwin, eds.), (Copenhagen, Denmark), pp. 94–106, Association for Computational Linguistics, Sept. 2017.

[88] A. Kumar, C. Tan, and A. Sharma, "Probing classifiers are unreliable for concept removal and detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17994–18008, 2022.

[89] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

# Publications from the Thesis

The following publications have resulted from the work presented in this thesis:

1. **Soumi Das**[*], **Shubhadip Nag**[*], Shreyyash Sharma, Suparna Bhattacharya, and Sourangshu Bhattacharya, "*VTruST: Controllable Value Function Based Subset Selection for Data-Centric Trustworthy AI*," in *Proceedings of the ICLR 2024 Workshop on Data-Centric Machine Learning Research (DMLR)*, International Conference on Learning Representations (ICLR), 2024. Available at: https://arxiv.org/abs/2403.05174. doi: https://doi.org/10.48550/arXiv.2403.05174.

2. **Shubhadip Nag**, Srinjoy Das, Agniva Saha, Anushree Ghosh, Soumi Das, Tarun Kumar, Suparna Bhattacharya, and Sourangshu Bhattacharya, "*MPSelectTune: Prompt-type Selection for Fine-tuning Improves Concept Unlearning in LLMs*," in *Proceedings of the NeurIPS 2025 Workshop on Reliable Machine Learning from Unreliable Data*, Conference on Neural Information Processing Systems (NeurIPS), 2025. Available at: https://openreview.net/forum?id=Jk8sOL97Tg.

[*]These authors contributed equally to this work.