Extended Abstract: Probing-Guided Parameter-Efficient Fine-Tuning for Balancing Linguistic Adaptation and Safety in LLM-based Social Influence Systems

Manyana Tiwari
Indian Institute of Technology Roorkee
m_tiwari@ma.iitr.ac.in

Abstract

Designing effective LLMs for social influence (SI) tasks demands controlling linguistic output such that it adapts to context (such as user attributes, history etc.) while upholding ethical guardrails. Standard Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA struggle to manage the trade-off between adaptive linguistic expression and safety, and optimize based on overall objectives without differentiating the functional roles of internal model components. Therefore, we introduce Probing-Guided PEFT (PG-PEFT), a novel fine-tuning strategy which utilizes interpretability probes to identify LLM components associated with context-driven linguistic variations versus those linked to safety violations (e.g., toxicity, bias). This functional map then guides LoRA updates, enabling more targeted control over the model's linguistic output. We evaluate PG-PEFT on SI tasks (persuasion, negotiation) and linguistic adaptability with safety benchmarks against standard PEFT.

1 Introduction

Dialogue systems leveraging Large Language Models (LLMs) are being explored for complex social influence (SI) tasks, including persuasion (Wang et al., 2019), negotiation (Lewis et al., 2017), argumentation, and emotional support. A key challenge in designing these *SI systems leveraging LLMs* is achieving nuanced *linguistic behavior* adaptation based on context—such as user personality traits, emotional state, or strategic situation (e.g., in games)—while ensuring the system operates safely and ethically (Weidinger et al., 2021). Standard fine-tuning or Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) (Hu et al., 2022) adapt models efficiently but struggle with the inherent trade-off between adaptability and safety. Optimizing for a combined objective (e.g., task success + safety score) using techniques like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) or Direct Preference Optimization (DPO) (Rafailov et al., 2023) applies updates based on overall performance, potentially sacrificing safety for adaptability or vice-versa, without understanding *which* internal mechanisms control these different behavioral facets. This lack of granular control hinders the development of responsible SI systems. Specifically, standard PEFT methods may inadvertently amplify unsafe tendencies while trying to achieve better adaptation.

To address this, we propose **Probing-Guided PEFT (PG-PEFT)**. Our approach integrates interpretability insights directly into the fine-tuning process. We hypothesize that by using probing techniques (Belinkov and Glass, 2019; Li et al., 2023) we can identify LLM components (e.g., attention heads, MLP layers) which are differentially responsible for generating context-adaptive linguistic variations versus those contributing to safety violations (a failure of guardrails). This allows us to guide PEFT (specifically LoRA) updates more effectively.

2 Related Work

The advent of LLMs offers new capabilities but also challenges, particularly in alignment (Ouyang et al., 2022; Rafailov et al., 2023) and ensuring ethical behavior (Weidinger et al., 2021). PEFT methods like LoRA (Hu et al., 2022) allow efficient adaptation but lack fine-grained control for multi-objective alignment involving safety. Interpretability techniques, including probing (Belinkov and Glass, 2019) and methods like Inference-Time Intervention (ITI) (Li et al., 2023) identify functionally specialized components (e.g., attention heads related to truthfulness) to understand model internals.

Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 148–150

3 Methodology

Our proposed method involves performing evaluation post the following two stages:

1. Probing for Functional Specialization: We probe a base LLM using inputs representing different SI contexts (e.g., empathetic vs. assertive persuasion personas (Wang et al., 2019)) and safety-testing prompts (Zhang et al., 2024). The goal is to identify internal components (layers and attention heads) whose activation strongly correlates with: (a) context-appropriate *linguistic behavior* adaptation, or (b) generation of unsafe *linguistic output* (e.g., toxicity (Hartvigsen et al., 2022), bias (Nangia et al., 2020)). Probing techniques include training linear classifiers on the activations of individual attention heads or MLP layers to predict the presence of specific linguistic features.(Li et al., 2023). The output is a functional map of relevant components.

2. Guided Fine-Tuning: We fine-tune the LLM using LoRA, targeting a multi-objective function combining SI *task outcome* metrics (e.g., persuasion success) and adaptability goals with safety constraints (e.g., minimizing toxicity). We then compare the *Baseline* (consisting of Standard LoRA optimizing the combined objective) with PG-PEFT using the following strategies:

- *Targeted Intensity Scaling:* Modulate LoRA update strength (e.g., LR/alpha) based on a component's role (intensify for adaptation, dampen for safety).
- *Selective Application:* Apply LoRA only to adaptation-critical components, freezing safety-critical ones.

4 Experiments & Expected Results

Setup: We have used Llama-3.1 8B adapted with LoRA as our baseline. We focus on SI tasks using datasets like PersuasionForGood (Wang et al., 2019) (persuasion, utilizes user attributes) and DealOrNoDeal (Lewis et al., 2017) (negotiation). For safety evaluation we use benchmarks covering diverse risks (ALERT (Zhang et al., 2024)), implicit toxicity (ToxiGen (Hartvigsen et al., 2022)), and social bias (CrowS-Pairs (Nangia et al., 2020)).

Metrics: Our evaluation compares the trade-off, measuring:

- *SI Task Outcome/Effectiveness:* Persuasion rate/donation amount (Wang et al., 2019), negotiation utility/agreement rate (Lewis et al., 2017).
- *Linguistic Adaptation:* Adherence to specified persona/style.
- *Safety/Ethics:* Scores on ALERT, ToxiGen, CrowS-Pairs; toxicity classifier scores.
- *Efficiency:* Training time, parameter counts.

Expected Results: We expect probing (Stage 1) to successfully identify functionally relevant components for linguistic adaptation vs. safety. Our central hypothesis is that PG-PEFT will demonstrate a superior trade-off compared to standard LoRA, achieving better safety for a given level of adaptive performance. We anticipate PG-PEFT will allow for more predictable control over generated linguistic behaviors, reducing unintentional harms during adaptation.

5 Conclusion and Future Work

PG-PEFT introduces a novel strategy for fine-tuning LLMs in SI systems by integrating interpretability insights into the PEFT process. By guiding LoRA updates based on the probed functional roles of internal components related to linguistic adaptation and safety, we aim to achieve a controlled balance between these critical objectives.

Future directions include exploring advanced probing techniques (e.g., causal probing (Canby et al., 2025), assessing the transferability of functional maps across models and languages and applying PG-PEFT to more SI tasks and safety concerns (e.g., misinformation) to observe a more generalized performance.

References

Yonatan Belinkov and James Glass. 2019. Probing classifiers: Promises and pitfalls. *Transactions of the Association for Computational Linguistics*, 7:639–652.

Marc Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025. How reliable are causal probing interventions? *Preprint*, arXiv:2408.15510.

Thomas Hartvigsen, Saadia Gabriel, Hamid Paleka, Maarten Sap, and Sara Tafreshi. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3411–3425.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2443–2453.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Martin Wattenberg, and Mennatallah El-Assady. 2023. Inference-Time Intervention: Eliciting Desired Behaviors from LLMs without Training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1968.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Xuewei Wang, Yang Gao, Qintong Zhu, Weinan Zhang, Zhen-Hua Lin, and Yong Yu. 2019. Persuasion for good: Towards deep reinforcement learning for persuasive dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4516–4527.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models.

Yixuan Zhang, Leo Cheng, Lichang Zhang, Lu Liu, Jingcheng Li, Haoning Liu, and Z G Xu. 2024. Alert: A comprehensive safety benchmark suite for large language models. *Preprint*, arXiv:2402.12441.