Research Statement

Moxin Li, Ph.D candidate at National University of Singapore

1 Introduction

Language models (LMs) and large language models (LLMs) have driven significant advances in AI, achieving top-tier performance across diverse language tasks. However, ensuring their trustworthiness, *i.e.,*whether humans can reliably depend on their behavior and outputs, remains a significant challenge, critical to responsible AI deployment and maximizing the societal benefit of LMs. The trustworthiness of LMs spans multiple dimensions. LMs should be robust to perturbations and sensitive to inputs with malicious intents. They should generate truthful responses, be honest to their limitations, avoid generating harmful content, ensure fairness, and preserve user privacy. Importantly, trustworthiness should be achieved without largely compromising general capabilities of LMs.

My PhD research addresses trustworthiness across three key stages of LM development (*cf.* Figure 1): during the **training stage**, enhancing **robustness** by mitigating spurious correlations in LMs and improving prompt optimization in LLMs; during the **inference stage**, enhancing **confidence calibration** by reducing overconfidence in LLM predictions; during the **alignment stage**, **balancing multi-dimensional trustworthiness** for LLM by resolving preference conflicts among different alignment objectives.

2 Enhancing Trust at Training Stage (2021 - 2023)

Reducing Spurious Correlations **[1]** This work proposes using hypothetical examples as a cost-effective method to reduce spurious correlations in fine-tuned LM for Machine Reading Comprehension (MRC) task, and a hypothetical training framework to encode causal relationships and largely improve robustness.

MRC is a crucial language understanding task. Fine-tuning LMs such as BERT has once been the predominant approach, where spurious correlations, also known as shortcuts, remain a major challenge on robustness. For instance, in table-based MRC, LMs may exploit shortcuts like over-relying on answer positions (*e.g.,* answers often appearing in the first column) to generate answer instead of true context understanding, leading to performance drops on examples without such shortcuts. Counterfactual training, *i.e.,*data augmenting with factual and counterfactual examples, has been a promising solution. However, the major concern is that building faithful counterfactuals is costly due to the complexity of preserving consistency and dependency in tabular data. To address this efficiently, we instead ask hypothetical questions, *e.g.,* "in which year would net profit be higher if 2019 revenue were $38,298?", which simulate the effect of counterfactual example without modifying tables. We propose a hypothetical training framework that uses paired examples with different hypothetical questions to supervise the direction of model gradient towards the counterfactual answer change. The superior results on tabular MRC datasets validate the effectiveness of our framework.

Robust Prompt Optimization **[2]** As prompt optimization became a highly promising research direction for black-box LLMs like ChatGPT, this work timely addressed the critical robustness challenge by proposing a gradient-free method that enhances generalization under domain shifts.

After the advent of ChatGPT, we have seen a paradigm shift from fine-tuning for specific tasks to prompting the state-of-the-art LLMs, making prompt optimization increasingly important. We revealed that the robustness issue remains a challenge in this scenario. The prompts are often optimized on labeled data from a specific distribution, yet the data LLM serves may have distribution shifts. We observe that applying optimized prompts on out-of-distribution data leads to significant performance drops. To address this, we formulate a new problem of robust prompt optimization for LLMs under distribution shifts, where prompts optimized on labeled source data should generalize to an unlabeled target group. We propose a generalized prompt optimization framework that integrates unlabeled target data into the optimization process by adapting knowledge from the labeled source group. Experiments show our framework significantly improves target group performance while maintaining comparable source group performance.

3 Enhancing Trust at Inference Stage (2023 - 2024)

Reducing Overconfidence **[3]** This work proposes an effective plug-and-play framework for reducing overconfidence in LLM self-evaluation, enhancing transparency and trustworthy of LLM responses.

Beyond robust performance, LLM should possess self-awareness of the reliability of its outputs. Specifically, LLM often generates outputs with factual errors. During inference, it is essential for LLMs to also express its confidence level that calibrates with the actual output correctness, thus supporting subsequent actions like abstention or human decision. However, LLM tends to inherently over-trust their incorrect outputs, assigning them high confidence scores and impairing confidence calibration. We believe this issue arises because existing self-evaluation methods only retrospectively evaluate outputs generated by LLMs. To address this limitation, we propose a novel self-evaluation paradigm that evaluates the comprehensive answer space beyond LLM-generated outputs, comparing the trustworthiness of multiple candidate answers to reduce over-trust in incorrect answers. Based on this, we introduce a two-step framework: first, LLMs reflect and justify each candidate answer; then, these justifications are aggregated for a thorough confidence estimation. This framework integrates easily with existing methods to enhance self-evaluation. Extensive experiments across six datasets and three tasks validate its effectiveness.

4 Enhancing Trust at Alignment Stage (2024 - 2025)

Balancing Multi-dimensional Trustworthiness **[4]** This work is the first to reveal and address the key issue of preference conflict in multi-objective alignment (MOA) with a self-improvement framework that enhances efficiency and scalability.

As LLMs become more powerful and trustworthy, the focus of trustworthiness should expand from single-dimensional to a multi-dimensional approaches. LLM outputs need to simultaneously achieve multiple desired features, such as harmlessness, helpfulness, factuality, and diversity, which can be achieved via multi-objective alignment (MOA). However, DPO-based MOA methods suffer from widespread preference conflicts, with different objectives favoring different responses. This results in conflicting optimization directions, hindering the Pareto Front optimization. To resolve this, we propose constructing Pareto-optimal responses to resolve preference conflicts, and introduce a self-improving DPO framework that enables LLMs to self-generate and select these responses for self-supervised preference alignment. Experiments on two datasets show our framework achieves a superior Pareto Front over baselines.

5 Future Directions

I believe achieving trustworthiness in LLMs requires moving beyond making them *perform like humans* to enabling them to truly *think like humans*. Several key principles including the pre-training, post-training, and test-time scaling laws have enable LLMs to reach human-level performance. However, the LLM generation does not subtly resemble human reasoning, and thus leads to issues such as non-robust, dishonest, and hallucination. I anticipate that trustworthiness will be achieved similarly as emergent abilities under next-step fundamental principles to largely align LLM toward more nuanced human-like thinking and behavior.

Therefore, I believe the next steps lie in both the ***evaluation stage*** and ***the reasoning stage***. In the evaluation stage, we need better evaluation methods for two key aspects: how well the LLM's generation align with the human thinking process, and whether the data used during training effectively supports this alignment. Current reward models, which largely focus on accuracy and general human preferences, are insufficient for capturing these deeper dimensions. We cannot enable LLM trustworthiness without a deep understanding on how to evaluate it. In the reasoning stage, recent advances in RL have significantly boosted LLMs' reasoning abilities, enabling it to explore the complex reasoning paths. Currently, the effort limits in well-defined tasks like mathematics and coding. In fact, humans tackle open-ended tasks through a latent, nuanced reasoning process, yet LLMs do not emulate such reasoning. Enhancing LLM reasoning combined with proper evaluation signal during the process is probably the path to fill this gap, which might not largely enhance performance but enhance trustworthiness.

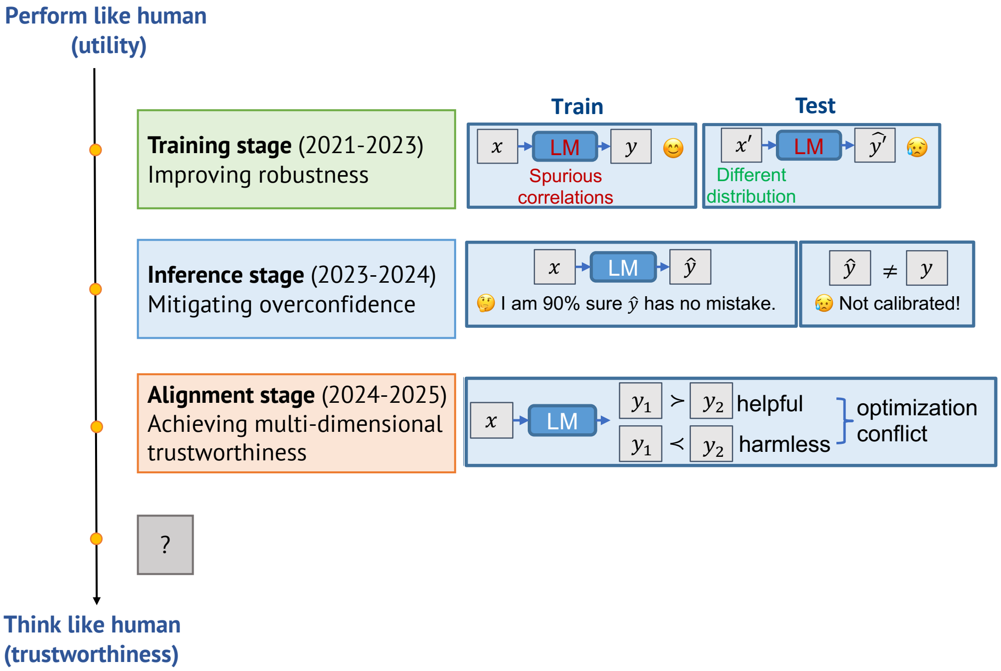

Figure 1: Research outline centering enhancing trustworthiness of LM.

References

[1] Moxin Li, Wenjie Wang, Fuli Feng, Hanwang Zhang, Qifan Wang, and Tat-Seng Chua. Hypothetical training for robust machine reading comprehension of tabular context. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1220–1236. Association for Computational Linguistics, 2023.

[2] Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. Robust prompt optimization for large language models against distribution shifts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1539–1554. Association for Computational Linguistics, 2023.

[3] Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11858–11875. Association for Computational Linguistics, 2024.

[4] Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. In *Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, July 27-August 1, 2025*. Association for Computational Linguistics, 2025.