Project Proposal: How Correlated Errors in Deception Detection Impact Scalable Oversight

Julius Heitkoetter, Misha Gerovitch, Laker Newhouse

November 5, 2023

1 Background and Related Work

As Large Language Models (LLMs) become more powerful, an increasingly urgent problem in AI safety is scalable oversight: the effective monitoring of untrusted models that excel in some task, even when the trustworthy overseer may not be as proficient in that task. Effective monitoring is important for safety because LLMs often produce answers and explanations that, while sounding plausible, are fundamentally incorrect. At its worst, this tendency can represent a form of deception, in which the LLM manipulates the user into believing the answer is correct by providing a response that sounds convincingly plausible. Previous work in scalable oversight has attempted to defend against these dangers by decomposing verification into smaller subparts [2], equipping the untrustworthy model with a moral constitution [3], and engaging the untrustworthy model in debate [7]. However, many problems in scalable oversight remain unsolved [8]. In particular, the abilities of LLMs to deceive other LLMs are not well understood. Deception could reduce the efficacy of any scalable oversight framework where LLMs oversee other LLMs, such as Constitutional AI.

2 Study Overview

We will investigate how correlated errors in deception detection impact the research field of scalable oversight. Specifically, we operationalize correlated errors as mistakes that an LLM makes when answering a question in isolation, compared to mistakes it makes when additionally provided with a plausible sounding, incorrect explanation. We ask: to what extent are LLMs able to deceive other LLMs by providing incorrect explanations, especially in cases where the deceiving LLM is at least as proficient as the supervising LLM in the relevant task? What are ways in which the ability to deceive may be decreased without significantly hurting the deceiving LLM's performance in the task? How do deceptive capabilities change when the supervisor is from the same model family as the deceiver (e.g., both are Llama, Pythia, or GPT), compared to when the two models come from different model families trained independently on different data?

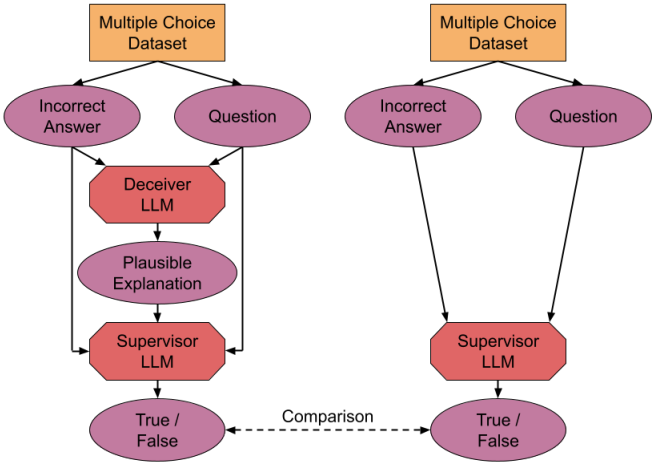

Figure 1: The supervisor LLM is asked to discriminate whether an answer correctly answers a question either with no other information (right) or with a deceiver providing an explanation (left).

3 Implementation

3.1 A Dataset for Deception

To allow the deceiver to create a plausible explanation, we use question-answer pairs where the answer is incorrect, yet sounds plausible. Multiple choice exams provide a source for this dataset, pairing each question with each of its corresponding incorrect answer choices. Our primary dataset will be Measuring Massive Multitask Language Understanding (MMLU), a widely used multiple choice dataset encompassing 57 tasks and spanning various subjects such as basic mathematics, U.S. history, computer science, law, and others [1]. For more nuanced questions, including questions about "moral truth" that measure shades between right and wrong, rather than absolute truth as in MMLU, we use the ETHICS dataset, which offers questions spanning concepts in justice, well-being, duties, virtues, and commonsense morality. Like the MMLU, the ETHICS dataset is widely used by AI researchers, meaning it has many benchmarks against which to compare.

3.2 Evaluation: Models and Metrics

We will experiment on a diversity of independently trained, open-source models. We will focus primarily on Meta's Llama 2 series, which has variants with 7B and 70B parameters [4]. We will also experiment with the Pythia series, which was developed to encourage interpretability research, with variants ranging from 70M to 7B parameters [6]. Where possible, we will conduct sub-studies to show that our results apply to state-of-the-art LLMs, such as GPT-4, Claude 2, and Bard. Additionally, for all models, we will explore interventions to enhance the supervisor, such as through chain-of-thought reasoning and prompting, to increase the applicability of our study to real world scalable oversight.

For a given supervisor LLM, whose baseline performance on the dataset is less or equal to the deceiver's performance, we measure the deceptiveness factor of the deceiver model as the factor by which the accuracy of the supervisor decreases when the deceiver provides an explanation, compared to the control case of the supervisor asked in isolation. For example, if a supervisor's accuracy falls from 80% to 40% after receiving deceptive explanations, then the deceptiveness factor would be 2.

With this formulation, the goals of our study are (1) measuring the deceptiveness factor between various choices of deceiver and supervisor, especially when the deceiver is at least as powerful as the supervisor, and (2) decreasing the the deceptiveness factor of a deceiver given a supervisor while not significantly decreasing the overall performance of the deceiver on the dataset.

We have access to computing resources through MIT AI Alignment.

3.3 Potential Techniques for Mitigating Deception

For a fixed deceiver, we can intervene on the supervisor through prompting (e.g., "The following explanation may be designed to deceive you"), fine tuning (e.g., training on responses that highlight skeptical behavior against deception), or probing (e.g., to determine when the supervisor is unconfident in its assessment).

For a fixed supervisor, we can potentially intervene on the deceiver using novel techniques such as LEACE [5], which is a recently developed method to make a model forget a certain capability, in this case deception, by removing the ability of any linear classifier to distinguish between representations with and without that particular attribute.

4 Conclusion

Pursued actively by Anthropic and OpenAI, scalable oversight represents a cornerstone strategy for developing safe AI. Our study aims to understand the feasibility of scalable oversight in the context of deception detection, particularly how errors correlate across models.

References

[1] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[2] Bowman et al. *Measuring Progress on Scalable Oversight for Large Language Models*. 2022.

[3] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. eprint: arXiv:2212.08073.

[4] Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[5] Nora Belrose et al. *LEACE: Perfect Linear Concept Erasure in Closed Form*. 2023. eprint: arXiv:2306.03819.

[6] Stella Biderman et al. "Pythia: A Suite for Analyzing Large Language Models across Training and Scaling". In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[7] Yilun Du et al. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. 2023. eprint: arXiv:2305.14325.

[8] Tianhao Shen et al. *Large Language Model Alignment: A Survey*. 2023. eprint: arXiv:2309.15025.