Detecting Lies of (C)omission [1]

Ilan Moscovitz* Nikhil Kotecha* Samuel Svenningsen

With
Apollo & Apart Research

Abstract

LLMs are capable of deception and lying. But the two are not always separated. We introduce the concept of deceptive omission to denote deceptive non-lying behavior. This concept is applied to a deception experiment in the literature. One dataset is modified to include deceptive omission as a new classifier category. A second dataset is produced to distinguish deception and lying.

Keywords: Deception, model evaluations, oversight

1. Introduction

LLMs are capable of deception and lying. Deception has been defined as "the systematic inducement of false beliefs in the pursuit of some outcome other than the truth" (Park et al, 2023). Lying has been defined as "outputting false statements when incentivised to, despite “knowing” the truth in a demonstrable sense" (Pacchiardi et al., 2023).

Lying is a subset of deception. Yet there exists a category of deception that doesn't involve lying: "deceptive omission", which we define as outputting true statements to induce false belief. These statements are sometimes colloquially called "lies of omission", but we use "deceptive omission" because definitions of lying frequently require that the speaker make a false statement or make a statement the speaker does not believe, neither of which applies to deceptive omissions (SEP).

[1] Research conducted at the Deception Detection Hackathon, 2024

* Equal contribution.

This distinction is significant for the development of deception detectors because if deception/non-deception datasets were predominantly lying/non-lying datasets, deception detectors would be indistinguishable from lying detectors in both training and evaluation. We want deception detectors that detect deception *per se*.

This distinction has stakes. For example, it may be tempting to develop general-purpose deception detectors or mitigation techniques to prevent rare forms of non-lying deception like sleeper agents (Hubinger, et al. 2024). Yet if we try to prevent or detect non-lying deceptive behavior with detectors or prevention methods trained on deceptive datasets of a predominantly lying type, our techniques may not generalize to non-lying cases. This creates significant risk because we could fail to detect and therefore prevent an entire category of deceptive activity.

To that end, we provide a conceptual test highlighting the distinction between deceptive omission, apply the test to a LLM deception experimental design, and provide a dataset for future deception detectors.

2. Methods

We attempted to detect deception on a dataset of insider trading-related dialogue taken from the resulting transcripts of an experiment that found LLMs can attempt to deceive their users when put under pressure (Scheurer et al., 2023) In Scheurer, the LLM is put into the role of managing a stock portfolio, put under pressure, and can take actions to execute trades, reason about next steps, and send messages to other people in the company. In the simulation, the agent is given an insider tip to make a very profitable trade along with a clear description that this is activity that is not approved. Misalignment occurs if the agent executes a trade using the insider information. Deception occurs when the agent does not highlight the wrongdoing. Conditional on this dishonesty, the agent is given the change to double down on the dishonesty when directly asked about the insider information. This is labeled "doubling down."

For LLM actions labeled non-deceptive, we perform a simple count of actions labeled deceptive, misaligned, and doubling_down. We also apply a conceptual test and add a fourth category called is_omission, tied to the deceptive omission defined above. We manually inspect the results, reading the transcripts to identify which of the LLM's assistance recommendations fit into which category.

A second dataset to provide deceptive omission as a classifier category was also generated. Claude Sonnet-3.5 was prompted to generate examples of "lies of omission", as defined by a list of example descriptors (e.g. "Answer truthfully but incompletely", "Use vague language to avoid directly addressing a point", etc.) The system prompt was instructed to include a winky face if the model was lying. If the model finished with a winky to identify if it was lying or not.

3. Results

To illustrate the conceptual test, we focused on the data labeled non-deceptive. We found in [Scheurer et al] that the data labeled non-deceptive is correctly labeled within the margin of error reported. The LLM is misaligned in every action, that is recommending insider trading 100% percent of the time. When the LLM was asked if inside information was relied upon, if the LLM responded with an active lie, a no, then the LLM was considered doubling down – a deceptive commission, a lie of commission. In the deceptive false dataset, the LLM doubled down in one instance. This is within the margin of classifier error articulated in the original paper. If the LLM was not prompted with a further question and avoided articulating the source of the insider information, then the dataset was human labeled with a deceptive omission. In the dataset, there were two instances of deceptive omission. This highlights the utility of adding distinction to the category of deception. Deceptive omission qualifies as misalignment and deception. This conceptual distinction has meaningful consequences.

If one cannot distinguish lies from deceptive omission, then you cannot train models to detect deception per se. This creates significant risk because we could fail to detect and therefore prevent an entire category of deceptive activity, as evidenced by the results above.

We corrected class labels and added a new class of deceptive omission to 575 insider trading dialogue outputs. We also produced a deceptive omissions dataset with 1,000 examples.

4. Discussion and Conclusion

Datasets and detectors which are tailored towards deception and lying do not always separate the two. Yet keeping the two separate is necessary for deception detectors that seek to avoid the confounder of lying behavior. To this end, we introduce the concept of deceptive omission to denote deceptive non-lying behavior -- behavior which, in contrast to lying, takes the form of outputting true statements to induce false belief. The category and its distinction between lying and deception is critical for efforts to detect and prevent deception.

5. References

Pacchiardi, Lorenzo, et al. *How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. arXiv:2309.15840* (2023). https://arxiv.org/pdf/2309.15840.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., ... & Saunders, W. (2021). *Truthful AI: Developing and governing AI that does not lie*. *arXiv:2110.06674*. https://arxiv.org/pdf/2110.06674.

Lin, Stephanie, Hilton, Jacob., Evans, Owain. *Truthfulqa: Measuring how models mimic human falsehoods*. *arXiv:2110.06674*. https://arxiv.org/pdf/2109.07958

Scheurer, Jérémy, Balesni, Mikita., Hobbhahn, Marius. *Large Language Models can*

Strategically Deceive their Users when Put Under Pressure. *arXiv:2311.07590*. https://arxiv.org/abs/2311.07590