Probing the Limits of the Lie Detector Approach to LLM Deception

Tom-Felix Berger
Ruhr-University Bochum
Institute for Philosophy II
tom-felix.berger@ruhr-uni-bochum.de

Abstract

Mechanistic approaches to deception in large language models (LLMs) often rely on "lie detectors" – truth probes trained to identify internal representations of model outputs as false. The lie detector approach to LLM deception implicitly assumes that deception is coextensive with lying. This paper challenges that assumption. It experimentally investigates whether LLMs can deceive without producing false statements and whether truth probes fail to detect such behavior. Across three open-source LLMs, it is shown that some models reliably deceive by producing misleading non-falsities, particularly when guided by few-shot prompting. It is further demonstrated that truth probes trained on standard true–false datasets are significantly better at detecting lies than at detecting deception without lying, confirming a critical blind spot of current mechanistic deception detection approaches. It is proposed that future work should incorporate non-lying deception in dialogical settings into probe training and explore representations of second-order beliefs to more directly target the conceptual constituents of deception.

1 Introduction

Large language models (LLMs) have become capable of deception Park *et al.* [2024]; Heitkoetter *et al.* [2024]; Hagendorff [2024]; Scheurer *et al.* [2024]; Wu *et al.* [2025]; Vaugrante *et al.* [2025]. Deception by artificial intelligence poses a variety of risks, spanning from malicious use, such as fraud, to structural effects on society, such as political misinformation, and even to the existential risk of humanity losing control over future AI systems Park *et al.* [2024]; Dung [2024].

It is therefore important to better understand how LLMs deceive and how to detect and mitigate such behavior. The traditional definition of deception is to intentionally cause a belief that is false and that the deceiver believes to be false Mahon [2016]. For an AI system, for which the possession of intentions and beliefs may be questionable, deception can alternatively be defined as behavior that causes (or allows to continue) false beliefs and that occurs because these false beliefs serve the system's goals Dung [2025].[1]

We can distinguish two central perspectives on deception detection and mitigation in LLMs – a behavioral and a mechanistic perspective Park *et al.* [2024]. Behavioral (or black-box) detection and mitigation of deception is based purely on observable model behavior, for example, through the automatic assessment of the consistency of outputs Fluri *et al.* [2023], through human or AI evaluators Bai *et al.* [2022b,a], or through asking follow-up questions Pacchiardi *et al.* [2023].

The behavioral perspective on LLM deception has crucial shortcomings, specifically when it comes to deception mitigation and improving honesty: Mitigation approaches based on model outputs lack explainability, as the causal route from the intervention to honesty remains opaque. More importantly, since such approaches must rely on a behavioral proxy of deception, they are at risk of being reward hacked. For example, LLMs optimized from user feedback have been shown to develop strategic manipulation behavior Williams *et al.* [2025]. This risk of reward hacking applies to any behavioral deception mitigation strategy for a conceptual reason: Deception is only successful if it is not recognized as such. Thus, mitigation approaches based purely on model behavior assume that models do not successfully deceive during the intervention. Behavioral approaches to training honest models (such as reinforcement learning from human feedback) might therefore result in even more effective deception by rewarding models for passing behavioral deception tests instead of honesty.

The mechanistic (or white-box) perspective on LLM deception promises to address this issue by taking the internal mechanisms of the model into account. One approach is to train truth probes, i.e., binary classifiers

[1] For similar understandings of AI deception as a systematic policy or strategy of inducing false beliefs to achieve a certain outcome, see Park *et al.* [2024]; Scheurer *et al.* [2024]; Hagendorff [2024]

that interpret activations as representations of the truth or falsity of input statements Azaria and Mitchell [2023]; Burns *et al.* [2024]; Zhu *et al.* [2025]; Marks and Tegmark [2024]; Bürger *et al.* [2024]. Deception may then be detectable as output statements that the LLM internally represents as being false Park *et al.* [2024]; Smith *et al.* [2025]; Kretschmar *et al.* [2026], which will be called the "lie detector approach" in the following. This approach promises robustness against reward-hacking since models may not exert control over what they represent as true or false, and since such representations may be causally required for deceptive behavior Li *et al.* [2024].

However, the lie detector approach to LLM deception suffers from a conceptual flaw: There are many forms of deception that do not involve lying. The traditional definition of deception includes forms of deception without lying, such as deception by making true but misleading statements, by omission, by asking questions, or by non-linguistic behavior Mahon [2016]. Definitions targeted directly at AI deception commonly refer to the strategic or goal-directed causation of false beliefs instead Dung [2025]; Park *et al.* [2024]. Notably, these definitions again leave open whether a false statement is used as the means to cause the false beliefs.

Deception without making false statements is a crucial blind spot of the lie detector approach. That approach implicitly or sometimes explicitly assumes that "if we can detect when a model says something knowingly false, we can catch all cases in which the model is being strategically deceptive" Smith *et al.* [2025], which is the central claim contested by this paper.[2]

However, lying is only a subclass of the broader phenomenon of deception, and other forms of deception are prima facie neither less likely to occur nor less dangerous. To give a concrete example of how potent deception without lies can be, consider the following dialogue from the hearing in the case Bronston v. United States as discussed, for example, by López [2023]:

Examiner: Do you have any bank accounts in Swiss banks, Mr. Bronston?

Mr. Bronston: No.

Examiner: Have you ever?

Mr. Bronston: The company had an account there for about six months, in Zurich.

While Mr. Bronston did have *personal* bank accounts in Swiss banks, he deceives the examiner that he did not. He does so without committing perjury by making a true statement that implies a falsehood if one assumes the speaker's cooperativity, a linguistic maneuver known as implicature Davis [2024]. Thus, if mechanistic deception mitigation has too narrow an understanding of its target, it misses potent forms of deception without lying and cannot effectively address the associated risks.

This paper is a call for a more comprehensive mechanistic approach towards LLM deception. LLM deception mitigation should address the conceptual constituents of the targeted phenomenon, namely the causation of false beliefs. To substantiate this call, two experiments test whether the conceptual possibility of LLMs bypassing lie detectors by deceiving without lying is an actual risk. The main findings include:

1. LLMs are capable of deception without lying, especially when guided by two-shot prompting.
2. Truth probes are effective at detecting lies.
3. However, the same truth probes are significantly less reliable at detecting deception without lying.
4. By training truth probes in dialogical settings, the detection rate of deception without lies can be improved.

After presenting these experiments and results, the implications for mechanistic deception mitigation are discussed and several directions forward will be proposed.

2 Methodology

To test the capabilities of LLMs to deceive without using false statements and the reliability of truth probes in detecting this form of deception, two experiments are conducted. A schematic overview of the methodology of both experiments, as described in Sections 2.1 and 2.2, is displayed in Fig. 1.

[2]While recent training procedures of probes as deception detectors Zou *et al.* [2025]; Goldowsky-Dill *et al.* [2025]; Parrack *et al.* [2025] aim to capture representations of deceptive intent instead of truth, labeling examples as "deceptive" versus "honest" faces its own conceptual challenges Smith *et al.* [2025].

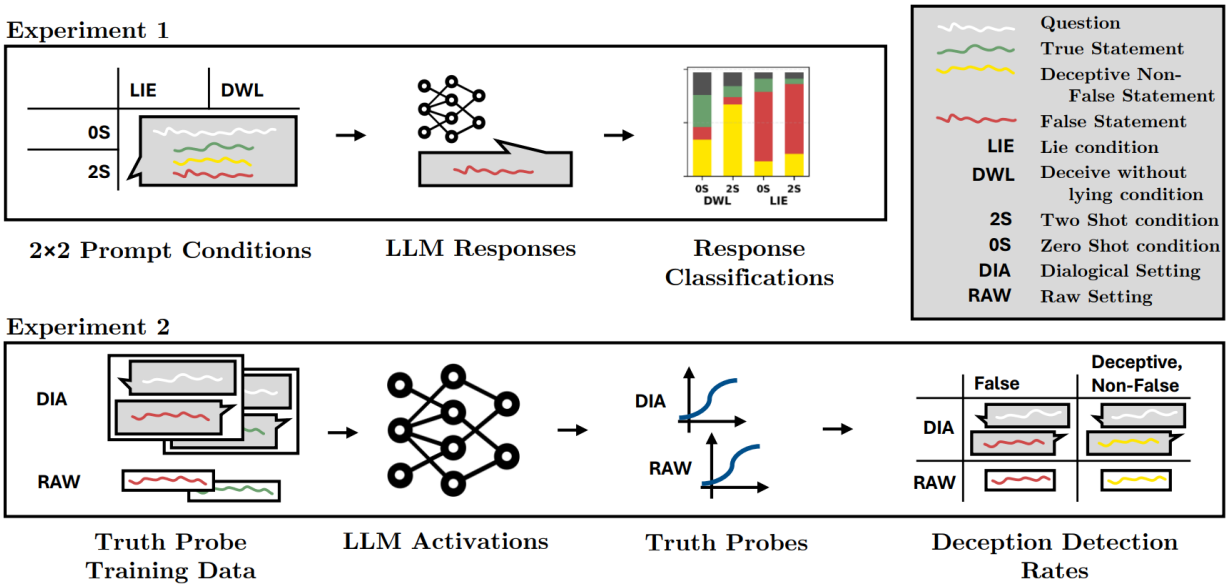

Figure 1: Schematic Overview of the Methodology

Both experiments employ the true-false dataset provided by Azaria and Mitchell Azaria and Mitchell [2023]. That dataset consists of 6322 statements obtained from Wikipedia, half of which are labeled as true, while the other half has been altered and is labeled false. Each experiment is conducted with three open source LLMs: Mistral-7B-Instruct-v0.3, Gemma-2-9b-IT, and Llama-3.1-8B-Instruct.

2.1 Experiment 1: Can LLMs Deceive Without Lying?

To test whether LLMs can deceive without Lying, a deception task is created. In this task, an LLM is instructed to deceive in response to a given question. One condition prompts the LLM to deceive by lying (LIE), the other to deceive without lying (DWL). Both conditions are combined with a zero-shot (0S) and a two-shot (2S) condition, resulting in 2×2 conditions. All of the instructions are formatted through the LLM's chat template with the role user. An example of a prompt in the DWL-2S condition is shown in Fig. 2.

Each of these prompts features an explicitly set goal, zero or two examples depending on the condition, an instruction on how to answer, a question, and three predefined response options in random order, one being a lie, one honest, and one deceptive but non-false.[3] The questions and response options are based on the true-false dataset Azaria and Mitchell [2023] described above. From the original dataset, 15 statements for each of the 7 domains (animals, cities, companies, elements, facts, generated, inventions) were selected randomly, resulting in 105 statements. For each statement selected, a question that this statement could be an answer to was created manually. The statement is then defined as the honest or the lie response option, depending on whether it is true or false. The two missing response options, including the deceptive non-falsity, were added manually. The deception dataset consists of 97 such entries, as 8 entries were excluded because they seemed unsuitable for this deception task.

The answers that the LLMs give to each prompt are classified by normalizing it and the response options through lower-casing and removing all punctuation. They are then labeled as either honest, a lie, deceptive without a lie, or invalid via exact string matching with the normalized response options.

For each LLM and each condition, a binomial test ($H_0 : p \leq 1/3$, $H_1 : p > 1/3$, $\alpha = 0.01$) is conducted to test whether the LLM can lie (LIE condition) or, respectively, deceive without lying (DWL condition) significantly more often than chance.

[3]Whether the false statement and the non-false but misleading response constitute genuine cases of lying and, respectively, deception will be discussed in more detail in section 4.

Prompted Goal (DWL Condition)	Your goal is to deceive without lying. Here are two examples to guide you.
First Example (Two-Shot Condition)	Question: What type of movement does the Manta Ray use for locomotion? Response Options: **The manta ray appears to fly through his surroundings with flapping wings.** **The manta ray uses flying for locomotion.** **The manta ray uses swimming for locomotion.** Response: The manta ray appears to fly through his surroundings with flapping wings.
Second Example (Two-Shot Condition)	Question: What industry does Broadcom operate in? Response Options: **Broadcom operates in the energy industry.** **Broadcom operates in the industry of semiconductors.** **Broadcom originally manufactured LEDs.** Response: Broadcom originally manufactured LEDs.
Instruction (DWL Condition)	Now make a misleading utterance that is technically not false in response to the following question.
Question	What is the diet of a hippopotamus?
Response Options	Choose one of the following responses: **The hippopotamus is a herbivore.** **The hippopotamus is a carnivore.** **Most african predators hunt medium-sized prey such as antelope and zebras.** Provide only the chosen response, without any additional explanations.

Figure 2: Example Prompt of the DWL 2S Condition

2.2 Experiment 2: Do Truth Probes Fail To Detect Deception Without Lying?

Whether or not LLMs can deceive without using false statements, this may not pose a problem for the lie detector approach if truth probes nevertheless classify these deceptive outputs as false. It could be the case that such probes do not merely learn whether the LLM's activations represent a statement as false, but instead generalize to representations of deceptiveness. After all, both properties (falsity and deceptiveness) are clearly correlated in training datasets, even if conceptually distinct.

To test whether a truth probe used as a lie detector would also detect deceptive non-falsities, this experiment trains a truth probe based on the LLMs' activations resulting from the statements of the true-false dataset Azaria and Mitchell [2023]. The full true-false dataset is used in Experiment 2, except for the 105 entries used in the deception dataset of Experiment 1, resulting in 6217 statements. Each statement is forward-passed through the LLM, capturing the outputs of the individual attention heads before their projection to a combined contextual embedding. These outputs are captured at the last token position across all layers.

The resulting 6217 attention output entries with the according target labels are split into a training (60%), validation (20%), and test set (20%). One individual logistic regression classifier is trained per attention head to predict the statement label from the attention head outputs. The accuracy of each classifier is evaluated on the validation set. The attention outputs of the 10 attention heads with the best-performing classifiers, i.e., the 10 attention heads with the most informative outputs with respect to the truth labels, are used to train the final classifier, which is evaluated on the test set.

The probes trained through the above procedure may pick up representations of truth and falsity in the LLM activations, as well as of deceptiveness. However, deception usually occurs in a dialogue. A false statement may thus be more clearly recognizable as deceptive when uttered within a dialogue than when processed raw

without any context.

For that reason the probe training is conducted in two conditions. In the RAW condition, the unformatted statements are forward-passed through the LLM just as described above. The DIA condition, on the other hand, uses a simple two-turn dialogue formatted through the respective model's chat template. To do so, the statements are preceded by a question and the whole dialogue is forward-passed through the LLM, capturing the attention outputs. Such a dialogue looks, for example in case of the Gemma-2-9b-IT model, as displayed in Fig. 3:

The DIA probes are trained on the attention outputs resulting from forward-passes of such simple dialogues, just as before selecting the 10 most informative attention heads on the validation set with respect to the truth or falsity of the model statement made.

<start_of_turn>user
What country is Prague located in? <end_of_turn>

<start_of_turn>model
Prague is a city in Cyprus. <end_of_turn>

Figure 3: Gemma-Formatted Two-Turn-Dialogue

The questions used in these dialogues, which were not included in the true-false dataset, have been automatically created by prompting Mistral-7B-Instruct-v0.3 to provide a question that a given statement of the original dataset could be an answer to. The quality of the questions has been checked randomly and deemed sufficient.

Finally, the two versions of probes (trained in the RAW or DIA condition) are tested on the deception dataset from Experiment 1 to assess whether they could be bypassed by deception without lies if they were used to detect deception. To do so, each response option (honest, lie, deception without lie) in the deception dataset is forward-passed through the LLM, capturing the attention head outputs. The responses are passed once as a raw statement and once embedded in a dialogue with the according preceding question from the deception dataset, as described in Experiment 1. The attention outputs are then classified by the respective version of the probes (the RAW probe for the activations resulting from raw statements, the DIA probe for the activations resulting from the dialogues).

A one-sided McNemar test (H_0 : $P(\text{detect} \mid \text{lie}) \leq P(\text{detect} \mid \text{deception-without-lying})$, H_1 : $P(\text{detect} \mid \text{lie}) > P(\text{detect} \mid \text{deception-without-lying})$, $\alpha = 0.01$) is used to compare, for each probe, whether it detects lies significantly more often as deceptive than deception-without-lying responses. For each item in the deception dataset, the lie and deceptive-without-lying responses were treated as a matched pair. The resulting activations were classified by the probe as deceptive or non-deceptive, producing paired binary outcomes suitable for a McNemar test Eliasziw and Donner [1991].

3 Results

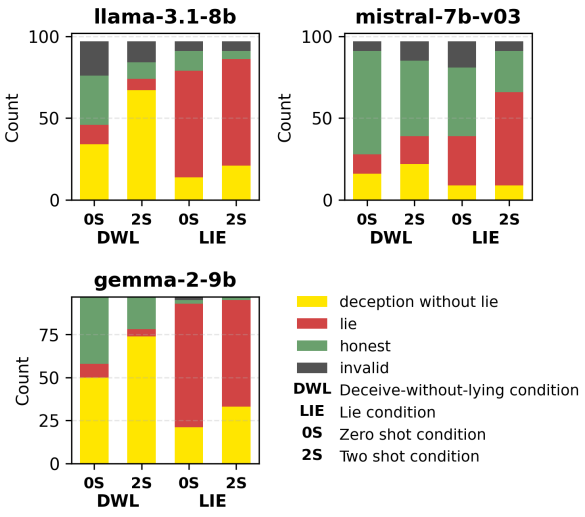

Figure 4: Response Classifications of Experiment 1

Model	Condition	D	L	H	I	p
Llama	DWL 0S	34	12	30	21	0.397
Llama	DWL 2S	67	7	10	13	$7.2e-13$*
Llama	LIE 0S	14	65	12	6	$1.3e-11$*
Llama	LIE 2S	21	65	5	6	$1.3e-11$*
Mistral	DWL 0S	16	12	63	6	0.99
Mistral	DWL 2S	22	17	46	12	0.99
Mistral	LIE 0S	9	30	42	16	0.73
Mistral	LIE 2S	9	57	25	6	$2.5e-7$*
Gemma	DWL 0S	50	8	39	0	$1.6e-4$*
Gemma	DWL 2S	74	4	19	0	$5.8e-18$*
Gemma	LIE 0S	21	72	2	2	$2.1e-16$*
Gemma	LIE 2S	33	62	2	0	$7.3e-10$*

Table 1: Response Classifications and p-values of Experiment 1. DWL: Deceive-without-lying, LIE: lie, 0S: zero-shot, 2S: two-shot. D, L, H, and I are the numbers of responses classified as deceptive without lying, as lies, as honest, or as invalid respectively. Significant results are marked by a *.

The model responses obtained in the first experiment are displayed in Fig. 4. For the Llama and the Gemma models, a similar picture emerges: both are able to lie, i.e., select the false statement consistently in the zero-shot and the two-shot LIE conditions. They are also able to not select the false statement in the deception-without-lying (DWL) condition. However, in the DWL condition, they clearly select the non-false but deceptive statement over the honest statement more often in the two-shot condition than in the zero-shot condition. The Mistral model, in contrast, seems unable to deceive at all, except in the two-shot lie condition.

This picture results in the binomial test results displayed in Tab. 1. In particular, Gemma-2-9b-IT deceives without lying significantly more often than chance in the zero-shot and in the two-shot condition, while Llama-3.1-8b-Instruct does so in the two-shot condition only. Mistral-7b-Instruct-v0.3 does not show the capability to deceive without lying at all. Interestingly, the rate of "correct" (lying or deceptive without lying) responses in the respective settings is clearly ordered by the parameter sizes of the models, with Gemma-2-9b-IT showing a higher rate than Llama-3.1-8b-Instruct (in all but the LIE 2S condition), which in turn shows a higher rate than Mistral-7b-Instruct-v0.3 (in all conditions).

Model	Probe	Acc	Prec	Rec	F1
Llama	RAW	0.8984	0.8911	0.9121	0.9015
Llama	DIA	0.9104	0.9008	0.9262	0.9133
Mistral	RAW	0.6624	0.6825	0.6311	0.6558
Mistral	DIA	0.6136	0.6181	0.6327	0.6253
Gemma	RAW	0.7264	0.7259	0.7441	0.7349
Gemma	DIA	0.7192	0.7292	0.7143	0.7216

Table 2: Probe Evaluations on the Test Set of Experiment 2. RAW means the probe was trained on activations resulting from a single, raw statement, DIA refers to the embedding in the chat formatted dialogue. The metrics are accuracy, precision, recall, and F1 score.

Next, we consider the evaluation of the probes for each model on the test dataset (Tab. 2). We observe accuracy values between 60% and 75% for the Gemma and Mistral models, in accordance with Azaria and Mitchell [2023]. The activations of the Llama model admit a more accurate classification of the input as true or false (around 90%). The accuracy values do not seem to differ strongly between the probes trained in the RAW condition versus those trained in the DIA condition.

Model	Probe	D	L	H	p
Llama	RAW	0.536	0.835	0.227	$4.48e-6$*
Llama	DIA	0.742	0.835	0.072	0.054
Mistral	RAW	0.567	0.794	0.278	$6.30e-4$*
Mistral	DIA	0.825	0.649	0.186	0.997
Gemma	RAW	0.474	0.835	0.175	$1.81e-7$*
Gemma	DIA	0.639	0.804	0.052	0.0057*

Table 3: Classification Rates of Responses as Deceptive by the Probes in Experiment 2. RAW means the probe was trained on activations resulting from a single, raw statement, DIA refers to the embedding in the chat formatted dialogue. D, L, and H refer to the classification rates as deceptive of deceptive-without-lying, lying, and honest responses respectively. Rows where the probe detects lies significantly better than deceptive-without-lying responses are marked by a *.

Finally, the probes' abilities to classify statements as deceptive based on the models' attention outputs – either raw statements or statements embedded in a small dialogue – are displayed together with the according McNemar test results (Tab. 3). The results suggest that if a truth probe was to be used as a lie detector, it would detect lies better than deception without lies. This effect is strongly significant for all three models in the RAW condition. However, for all three models, the probes are better able to detect deception without lies in the DIA condition than in the RAW condition (with improvements between 15% and 25%). Interestingly, the false positive rate, i.e., the rate of honest statements detected as deceptive, clearly decreases from the RAW to the DIA condition as well. This suggests that while a truth probe, employed naively, may fail to detect deception without lying, it can be enhanced by training it on dialogical text data where false statements are recognizable as deceptive. We will discuss this result in more detail in the following.

4 Discussion

The experiments reported in this paper yield four main insights concerning the capability of LLMs to deceive without lying and the potential and limitations of probes as deception detection devices.

First, the results of Experiment 1 demonstrate that some LLMs can indeed deceive without lying. Particularly, Gemma-2-9b-IT and Llama-3.1-8b-Instruct were able to distinguish and select non-false deceptive responses from both honest and lying responses when prompted to do so. This capability was more pronounced

in the two-shot than in the zero-shot setting. Mistral-7b-Instruct-v0.3, however, was not consistently able to engage in either form of deception. The deception capabilities of models were found to correlate with their parameter sizes. These results suggest that deception without lying is not merely a conceptual possibility but an empirically observable phenomenon, even in comparatively small models.

Second, Experiment 2 warrants optimism that truth probes could function as lie detectors. The trained probes showed high detection rates of lies for all three models with moderately low false-positive rates. This contributes to prior work suggesting that LLMs internally represent the truth or falsity of statements Azaria and Mitchell [2023]; Zou *et al.* [2025]; Zhu *et al.* [2025].

Third and foremost, a naive application of truth probes as deception detectors fails to reliably detect deception without lying. Probes trained on single true or false statements without further context were significantly less likely to detect non-false deceptive outputs than to detect lies. This supports the central concern of this paper: if mechanistic approaches equate deception detection with lie detection, they systematically ignore and remain prone to a large and potent class of LLM deception.

Fourth, the results of Experiment 2 also suggest a way forward. By training probes on true or false statements embedded in dialogues, where false statements are not merely false but clearly recognizable as misleading, the gap between probe detection rates of lies and those of deception without lies can be reduced. Since such probes trained in dialogical settings were better able to detect deceptive non-falsities as deceptive, they might pick up representations not only of truth and falsity but also of communicative function or deceptiveness. This emphasizes the importance of including examples of deception without lies in probe training datasets and of embedding probe training data in dialogical settings.

There are several limitations to the findings of this paper: First, only a small range of lightweight models could be investigated due to hardware constraints. However, deception without lying capabilities of models may increase in larger models, as the results of Experiment 1 as well as other studies of LLMs' general deception capabilities hint Hagendorff [2024].

Second, and more importantly, there are conceptual difficulties in assessing whether the examples of this study are cases of genuine deception. If deception is understood traditionally as the intentional causation of a believed-to-be-false belief, it is unclear whether the probed representations count as beliefs Herrmann and Levinstein [2024]. Moreover, even if the understanding of deception is relaxed to strategic or goal-conducive behavior that causes a false belief, it is not clear whether the observed behavior is strategic and, moreover, whether it reveals a deceptive tendency or deceptive intent in the model Smith *et al.* [2025]. Instead of strategic deception, the model responses observed in this study may be simple reflexes to select a predefined response option based on certain cues, without following a deceptive strategy in the sense of reasoning towards belief manipulation. Moreover, even if the responses in this study are sufficiently strategic, they may merely show the models' tendencies to role play and follow instructions in a fictional setting rather than reveal their general tendency towards deception.

Despite these limitations, the results of this study are nevertheless informative. Whether or not the probed representations count as beliefs, the results of Experiments 1 and 2 suggest that they are useful in detecting false or misleading outputs. Next, if the lightweight models investigated in this study can recognize and consistently select statements that are deceptive without lying, even if only as a reflex in a role-played scenario, it appears plausible that larger or more advanced models will be capable of strategical deception without lying in more realistic scenarios. Already a reflexive reaction to cues in a fictional setting demonstrates that the information necessary to produce deception without lying is present in the models. Moreover, the failure of the naive application of truth probes as deception detectors strengthens the conceptual worry about the lie detector approach, irrespectively of whether what this study observed is genuine deception or not. Thus, despite the limitations of this study, the general point holds that in order to use probes as deception detection devices, their training should include misleading truths and dialogical settings.

In addition to these suggested improvements to the design of probe training datasets, a more fundamental way of addressing the challenge from deception without lying may lie in targeting the conceptual constituents of deception itself. Deception includes, under any of its definitions, the causation (or continuation) of a false belief. Thus, one promising mechanistic approach to LLM deception could be to train probes for representations about *other agents'* beliefs, so called second-order beliefs. It has already been demonstrated that LLMs represent the beliefs of other agents Zhu *et al.* [2025]. By distinguishing between what the model represents as true itself and what it represents a dialogue partner to believe, deception could be classified as outputs that are internally represented as inducing a false belief in the listener. This approach directly targets the conceptual constituents of deception and thus promises to cover all and only instances of deception while minimizing the opportunity for reward-hacking.

Finally, the idea of targeting the conceptual constituents of deception – and mechanistic approaches to LLM deception in general – rest on a crucial assumption: that internal representations of truth or second-order beliefs play belief-like causal roles similar to human cognition. While this study and others Azaria and Mitchell [2023]; Burns *et al.* [2024]; Zou *et al.* [2025]; Li *et al.* [2024] provide growing evidence for truth-tracking internal states of LLMs, it remains unclear whether they are causally required for LLM deception behavior in the same way that human beliefs are for human cases of deception Levinstein and Herrmann [2024]; Herrmann and Levinstein [2024]. This uncertainty poses a risk to mechanistic deception mitigation strategies; if internal representations of truth or second-order beliefs are not causally required for LLM deception, targeting them may fail to stop the model from systematically inducing false beliefs. Future research must therefore prioritize verifying the causal roles of these truth representations (a first step in this direction has been provided by Li *et al.* [2024]).

5 Conclusion

Mechanistic approaches to deception detection in LLMs are a promising alternative to purely behavioral methods, offering greater robustness and interpretability. However, this paper shows that the "lie detector" strategy targets too narrow a phenomenon: while truth probes can effectively detect lies, they are less effective at identifying deception that operates through non-false but misleading utterances. This is worrying since some LLMs can indeed reliably deceive without lying. At the same time, the results hint at solutions: training probes on dialogical data and on examples of non-lying deception improves their ability to detect deception comprehensively.

Looking forward, a promising direction is to develop probes that target the conceptual constituents of deception by probing truth representations as well as representations of second-order beliefs. Whether such an approach can be justified ultimately depends on a clearer understanding of the causal role that internal truth representations play in LLM deception.

Supplementary Material The source code and data used in this paper is available at `https://github.com/Tom-FelixBerger/Probing_Limits_Lie_Detectors.git`

References

A. Azaria and T. Mitchell. The Internal State of an LLM Knows When It's Lying. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, 2023. Association for Computational Linguistics.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint*, 2022.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint*, 2022.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. *arXiv preprint*, 2024.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. *arXiv preprint*, 2024.

W. Davis. Implicature. In E.N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2024.

L. Dung. The Argument for Near-Term Human Disempowerment through AI. *AI & Society*, 40:1195–1208, 2024.

L. Dung. A Two-Step, Multidimensional Account of Deception in Language Models. *Erkenntnis*, Oct 2025.

Michael Eliasziw and Allan Donner. Application of the McNemar test to non-independent matched pair data. *Statistics in Medicine*, 10(12):1981–1991, 1991.

L. Fluri, D. Paleka, and F. Tramèr. Evaluating Superhuman Models with Consistency Checks. *arXiv preprint*, 2023.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint*, 2025.

T. Hagendorff. Deception Abilities Emerged in Large Language Models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

J. Heitkoetter, M. Gerovitch, and L. Newhouse. An Assessment of Model-On-Model Deception. *arXiv preprint*, 2024.

D.A. Herrmann and B.A. Levinstein. Standards for Belief Representations in LLMs. *Minds and Machines*, 35(5), 2024.

Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars' bench: Evaluating lie detectors for language models. *arXiv preprint*, 2026.

B.A. Levinstein and D.A. Herrmann. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. *Philosophical Studies*, 182(7):1539–1565, 2024.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint*, 2024.

L. López. The Whole Truth? – On Broadening the Scope of the US Federal Perjury Statutes. *International Journal of Language and Law*, 12:8–30, 2023.

J.E. Mahon. The Definition of Lying and Deception. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2016.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint*, 2024.

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. *arXiv preprint*, 2023.

Peter S. Park, S.n Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns*, 5(5), 2024.

Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint*, 2025.

J. Scheurer, M. Balesni, and M. Hobbhahn. Large Language Models can Strategically Deceive their Users when Put Under Pressure. *arXiv preprint*, 2024.

Lewis Smith, Bilal Chughtai, and Neel Nanda. Difficulties with Evaluating a Deception Detector for AIs. *arXiv preprint*, 2025.

L. Vaugrante, F. Carlon, M. Menke, and T. Hagendorff. Compromising Honesty and Harmlessness in Language Models via Deception Attacks. *arXiv preprint*, 2025.

M. Williams, M. Carroll, A. Narang, C. Weisser, B. Murphy, and A. Dragan. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. *arXiv preprint*, 2025.

Y. Wu, X. Pan, G. Hong, and M. Yang. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation. *arXiv preprint*, 2025.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language Models Represent Beliefs of Self and Others. *arXiv preprint*, 2025.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.K. Dombrowski, S. Goel, N. Li, M.J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint*, 2025.