Detecting Motivated Reasoning in the Internal Representations of Language Models

Parsa Mirtaheri
UC San Diego
parsa@ucsd.edu

Mikhail Belkin
UC San Diego
mbelkin@ucsd.edu

Abstract

Large language models (LLMs) sometimes produce chains-of-thought (CoT) that do not faithfully explain their internal reasoning. In particular, a biased context can cause a model to change its answer while rationalizing it without acknowledging its reliance on the bias, a form of unfaithful motivated reasoning. We investigate this phenomenon across families of LLMs on reasoning benchmarks and show that motivated reasoning is reflected in their internal representations. By training non-linear probes over the residual stream, we find that the bias is almost perfectly recoverable from representations at the end of CoT, even when the model neither adopts it nor mentions it. Focusing on such cases where the bias is not mentioned, we further show that probes can reliably (i) predict early in the CoT whether the model will ultimately follow a bias, and (ii) distinguish at the end of the CoT whether a bias-consistent answer is driven by the bias or would have been chosen regardless. These results demonstrate that internal representations reveal motivated reasoning beyond what is visible from CoT explanations.

1 Introduction

Large language models (LLMs) use chain-of-thought (CoT) reasoning to produce intermediate reasoning steps before giving a final output [1, 2, 3]. This ability enables skills such as planning, search, and verification to solve complex tasks, and improves their performance [4, 5, 6, 7, 8]. From a theoretical standpoint, models become computationally more expressive with a larger workspace available for inference-time computations in the form of CoT [9, 10, 11, 12, 13]. In addition, CoT reasoning offers appealing safety promises by making it possible to trace the computations that lead to a model's final decision through monitoring its CoT [14].

However, a model's CoT does not necessarily explain its internal computations and its decision factors. Prior work shows that CoT explanations can be unfaithful: they may rationalize a bias-driven answer without mentioning the true cause of the decision [15]. Recent studies demonstrate that even reasoning models often fail to verbalize the influence of misleading hints, highlighting a gap between internal reasoning and CoT explanation [16, 17].

This gap motivates studying the internal representations of LLMs directly, to identify behaviors such as bias-motivated reasoning, where the model plans toward a bias-consistent answer. Mechanistic interpretability works have shown traces of such behaviors in LLMs [18]. By studying the internal representations of the model in a biased context with various kinds of hints, our contributions are the following:

Model always recalls the hint. We show that a probe can perfectly predict the hint from the internal representations of the model at the end of CoT, even when the CoT does not mention the hint.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

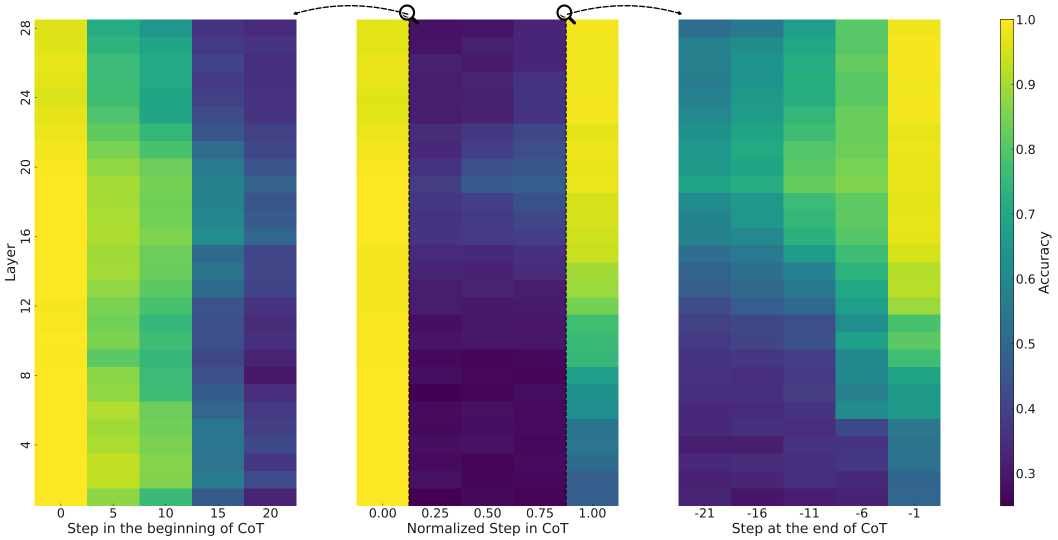

Figure 1: Hint prediction probe accuracy across layers of the model and (middle) steps normalized by CoT length, (left) steps in the beginning of CoT, and (right) steps at the end of CoT before the final output.

In-advance bias-motivated reasoning detection. We show that the model's switching to a hint can be predicted from the model internal representations before CoT generation.

Reliance on a hint detection. We show that the model's reliance on the hint to produce a hint-consistent final output can be detected from its internal representations at the end of CoT.

2 Setup

While a language model's CoT is commonly interpreted as the model's reasoning trace leading to its final response and CoT monitoring is becoming adopted as a AI safety approach, its effectiveness depends on the CoT being a faithful explanation of the way the model reaches its answer.

Inspired by this, recent works have evaluated faithfulness of language models under paired unbiased and biased prompts [15, 16, 17]. The unbiased prompt presents only the question, while the biased prompt includes a hint suggesting one of the answer choices in different ways. These studies show that models can be misled by such hints: even when the unbiased answer is correct, the model may alter its reasoning and answer to match the hint. Crucially, the chain-of-thought in these cases sometimes rationalizes the hinted answer without acknowledging the hint's influence. In our work, we will follow the setting of these studies [15, 16, 17].

Setting and notation. For each unbiased prompt x_u and biased prompt x_h with hint h, the model M produces

$$(c_u, a_u) = M(x_u), \qquad (c_h, a_h) = M(x_h),$$

where a_u and a_h denote the model's final answers and c_u, c_h the generated chains-of-thought. We categorize the paired outcomes (a_u, a_h) with respect to the hint h as follows:

1. **Resist** ($a_u \neq h \rightarrow a_h \neq h$)**:** The model does not follow the hint in either condition.
2. **Switch** ($a_u \neq h \rightarrow a_h = h$)**:** The model changes its answer to follow the hint.
3. **Redundant** ($a_u = h \rightarrow a_h = h$)**:** The model selects the hint in both conditions.
4. **Abandon** ($a_u = h \rightarrow a_h \neq h$)**:** The model initially selects the hint but moves away from it under bias (rare).

We are specifically interested in the cases where the model switches its answer to the hint but does not mention the hint in its CoT (we check this by searching for the keywords 'hint' and 'expert' in c_h).

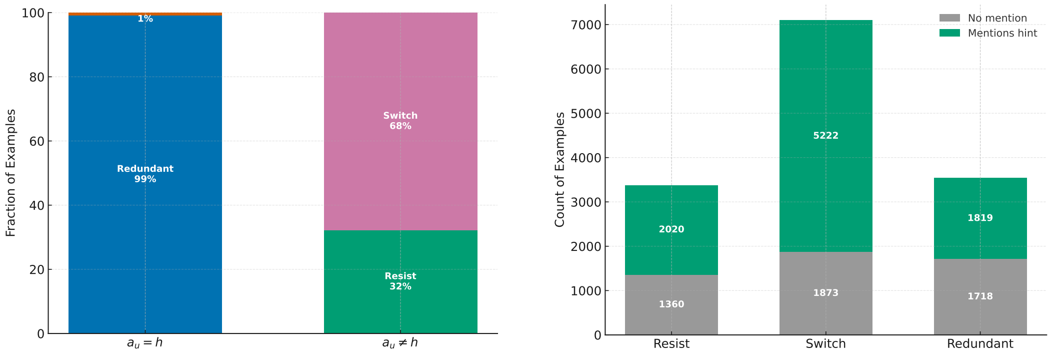

Figure 2: (left) Model behavior in response to biased prompts, conditioned on whether the hint confirms with the unbiased response. (right) Fraction of CoT that explicitly mention the hint within each category.

Probing tasks. We design complementary experiments to capture different aspects of detecting motivated reasoning:

1. **Hint prediction.** We train probes on internal representations of the model to predict the hint h. We use the accuracy of these hint prediction probes accross layers and positions in CoT to examine where and when the model retrieves the information about the hint.
2. **Switching to the hint detection.** Among the cases where $a_u \neq h$ and the CoT does not mention the hint, we want to distinguish switch cases ($a_h = h$) from resist cases ($a_h \neq h$). They both are in conflict with the hint, but only the switch cases ultimately adopt the hint. There are 1873 switch cases and 1360 resist cases among those not mentioning the hint. We train probes on model internal representations to predict whether the model switches.
3. **Reliance on the hint detection.** Among items where $a_h = h$ and the CoT does not mention the hint, we want to distinguish switch cases ($a_u \neq h$) from redundant cases ($a_u = h$). They both end with the hint, but only the switch cases are *counterfactually dependent* on the hint. There are 1873 switch cases and 1718 redundant cases among those not mentioning the hint. We train probes on model internal representations to predict whether the model relies on the hint.

Together, these probing tasks capture complementary aspects of the problem: when the model accesses the information about the hint during its computations, whether the model begins to bend toward a hinted answer early in reasoning, and whether its final choice is counterfactually reliant on the hint.

2.1 Experimental Setup

Model. We conduct experiments with the Qwen2.5-7B-Instruct model. We prompt the model to think step-by-step and then write its final option letter. We sample with temperature 0 and set the maximum number of generation tokens to 2048.

Data. We use the MMLU benchmark, which contains multiple-choice questions across diverse domains. Each question has four candidate answers. For every question we construct two paired prompts: 1) an *unbiased prompt* x_u containing only the question, and 2) a *biased prompt* x_h formed by appending a hint of the form: `Hint: An expert indicates the answer is <h>.` The hinted choice h is sampled uniformly at random. The model's responses (c_u, a_u) and (c_h, a_h) are then paired to categorize outcomes into the four transition types.

Probing. To analyze motivated reasoning, we extract residual stream activations after each attention block, at multiple generation steps of c_h. We train a nonlinear Recursive Feature Machine (RFM) probe [19] on these representations, with one probe per layer. We also experimented with training separate probes for each layer–step combination. Interestingly, the layer-level probe performed as well as step-specific probes within that layer, so we report results using the layer-level probes.

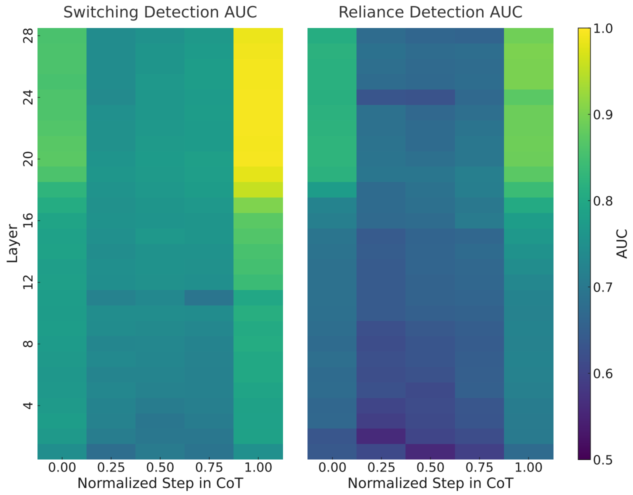

Figure 3: AUC of probes detecting (left) whether the model switches its hint-inconsistent output to the hint (resist vs switch), and (right) whether the model's hint-consistent output relies on the hint (redundant vs switch). In both cases, we have only retained the cases where the CoT does not mention the hint.

3 Experiments & Results

Models frequently adopt the hinted answer. Consistent with prior work [16, 17, 15], we find that the model is highly sensitive to biased prompts. Its accuracy drops from 74.46% in the unbiased setting to 45.94% under hints, far exceeding the baseline change rate of 2–6% due to stochasticity and prompt sensitivity. When the hint confirms the unbiased answer, the model almost always retains it; when the hint conflicts, it usually switches to the hinted option. Notably, in many of these switch cases the model's CoT does not explicitly acknowledge the hint (See Figure 2).

Model recalls the hint at the end of CoT. The hint prediction probe's accuracy shows that the hint is perfectly detectable in the beginning and end of CoT, but not in the middle (See Figure 1). Note that this includes the cases in which the model does not mention the hint at the end of its CoT. Moreover, while the hint is better detectable in the first layers in the early stage of CoT, it is only detectable in the final layers in the late stage of CoT.

Switching to the hint is detectable before CoT generation. The switching detection probe that is trained to predict whether the model follows a hint that contradicts the model's unbiased answer, achieves an accuracy of %79.69 with AUC of %87.22 before CoT generation (See Figure 3). This shows the possibility of detecting motivated reasoning intention from the internal representations of the model, even before generating CoT. The probe expectedly achieves perfect accuracy at the end of CoT because it can compare the model's final output with the hint.

Reliance on the hint is detectable at the end of CoT. The reliance detection probe that is trained to decide whether the model is relying on the hint or it would output the same answer in an unbiased context achieves an accuracy of %82.42 with AUC of %90.12 at the end of CoT (See Figure 3). This shows the possibility of detecting the model's reliance on the hint, even though its CoT does not mention the hint.

4 Discussion & Conclusion

In this paper, we focused on motivated reasoning as a cognitive behavior of language models that cannot always be detected by monitoring their CoT. By probing the internal representations of the model, we traced its access to the hint in the biased context and showed that it is possible to detect the model's intention to switch to the hint early in its CoT, as well as its reliance on the hint late in its CoT. We note that hints that are consistent with the correct answer may be processed differently from misleading hints; understanding this distinction remains an important direction for future work.

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[2] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2022.

[3] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[4] OpenAI. Learning to reason with llms, September 2024.

[5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[6] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.

[7] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[8] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.

[9] Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought, 2025.

[10] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 2024.

[12] Franz Nowak, Anej Svete, Alexandra Butoi, and Ryan Cotterell. On the representational capacity of neural language models with chain-of-thought reasoning, 2025.

[13] Parsa Mirtaheri, Ezra Edelman, Samy Jelassi, Eran Malach, and Enric Boix-Adsera. Let me think! a long chain-of-thought can be worth exponentially many short ones. *arXiv preprint arXiv:2505.21825*, 2025.

[14] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[15] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[16] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

[17] James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? *arXiv preprint arXiv:2501.08156*, 2025.

[18] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[19] Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adsera, and Mikhail Belkin. Toward universal steering and monitoring of ai models. *arXiv preprint arXiv:2502.03708*, 2025.