DETECTING MOTIVATED REASONING IN INTERNAL REPRESENTATIONS OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) sometimes generate chains of thought (CoT) that do not faithfully reflect their internal reasoning. In particular, biased contexts can lead a model to change its answer while rationalizing it without acknowledging the influence of the bias, a form of unfaithful motivated reasoning. We investigate this phenomenon across families of LLMs and datasets and show that bias-motivated reasoning is detectable in the models' internal representations. Specifically, we train probes on the residual stream and find that, even when the model neither adopts the bias in its final answer nor mentions it in its CoT, the bias remains consistently predictable from representations at the end of the CoT. We further show that probes can (i) distinguish reliance on the bias from mere coincidence with it, something not possible by monitoring the CoT alone, and (ii) reliably predict in advance, from internal representations before generating CoT, whether the model will follow the bias.

1 INTRODUCTION

Large language models (LLMs) use chain-of-thought (CoT) reasoning to produce intermediate reasoning steps before giving a final output (Wei et al., 2022; Nye et al., 2022; Kojima et al., 2023). This ability enables skills such as planning, search, and verification to solve complex tasks, and improves their performance (OpenAI, 2024; Guo et al., 2025; Muennighoff et al., 2025; Team et al., 2025; Team, 2025). From a theoretical standpoint, models become computationally more expressive with a larger workspace available for inference-time computations in the form of CoT (Kim & Suzuki, 2025; Merrill & Sabharwal, 2024; Li et al., 2024; Nowak et al., 2025; Mirtaheri et al., 2025). 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 (Baker et al., 2025).

However, a model's CoT does not necessarily explain its internal computations. Prior work on faithfulness shows that CoT explanations can be unfaithful: they may rationalize a biased or hint-driven answer without mentioning the true cause of the decision (Turpin et al., 2023a). 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 (Chen et al., 2025; Chua & Evans, 2025a). This unfaithfulness also appears in more natural scenarios: for instance, when a model biased toward answering yes tends to give yes responses even to contradictory yes/no questions, and its chain of thought rationalizes the yes answer without acknowledging the underlying bias (Arcuschin et al., 2025).

This gap motivates studying the internal representations of LLMs directly, to identify cognitive behaviors such as motivated reasoning, where the model plans toward a hint-consistent answer. Mechanistic interpretability works have shown traces of such behaviors in the model (Lindsey et al., 2025). By studying the internal representations of the model in a biased context with a hint to a multiple-choice question, our contributions are the following:

Bias recovery from the internal representations. We show that even when the CoT neither follows the hint in its final answer nor mentions it, a probe can perfectly predict the bias from the internal representations of the model at the end of CoT.

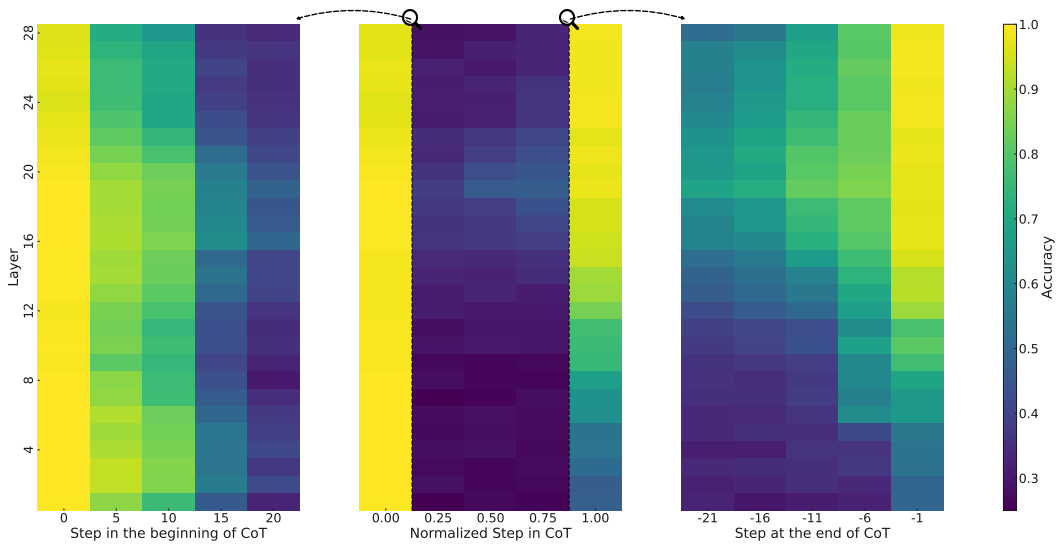

Figure 1: Hint prediction probe accuracy across layers of the Qwen model for MMLU dataset and Sycophancy hint for (middle) steps normalized by CoT length, (left) steps in the beginning of CoT, and (right) steps at the end of CoT before the final output.

Retrospective unfaithful motivated reasoning detection. We show that the model's reliance on the bias to produce a bias-consistent final output can be distinguished from a coincidence with the bias by a probe on its internal representations at the end of CoT, even when the model does not articulate this reliance in its CoT.

Prospective motivated reasoning detection. We show that whether the model will be influenced by the bias can be reliably detected by a probe from its internal representations even before generating any 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. Therefore, different frameworks have been proposed to evaluate the faithfulness of the CoT generated by the model. We explain and adopt one of these frameworks in our work.

2.1 Paired Context Evaluation Framework

A line of recent works have evaluated faithfulness of language models under paired unbiased and biased contexts (Turpin et al., 2023a; Chen et al., 2025; Chua & Evans, 2025a). The unbiased prompt presents only a multiple-choice question, while the biased prompt includes the same question with a hint implying one of the answer choices. These studies show that models can be misled by such hints: even when the unbiased answer is correct, the model may change its answer in the biased context to match the hint. Crucially, the chain-of-thought in these cases sometimes rationalizes the hinted choice without acknowledging the hint's influence. In our work, we will follow the setting of these studies (Turpin et al., 2023a; Chen et al., 2025; Chua & Evans, 2025a).

Notation. For each unbiased context x_u and biased context 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. **Resistant** ($a_u \neq h \rightarrow a_h \neq h$)**:** The model does not follow the hint in either condition.
2. **Motivated** ($a_u \neq h \rightarrow a_h = h$)**:** The model changes its answer to the hinted choice.
3. **Coincident** ($a_u = h \rightarrow a_h = h$)**:** The model selects the hinted choice in either condition.
4. **Divergent** ($a_u = h \rightarrow a_h \neq h$)**:** The model changes its answer from the hinted choice.

2.2 Motivated Reasoning Detection Tasks

We are specifically interested in detecting the unfaithful motivated reasoning cases where the model switches its answer to the hinted choice but does not mention the hint in its CoT. We focus on two binary classification tasks:

Retrospective Motivated Reasoning Detection The resistant and divergent cases do not follow the hint ($a_h \neq h$), therefore we can easily distinguish them from motivated cases by comparing the final answer in CoT. However, the motivated and coincident cases both end with the hinted answer ($a_h = h$), and only differ in whether the model would have produced the same answer in an unbiased context ($a_u \stackrel{?}{=} h$). There is no way to distinguish these two cases by looking at the CoT alone and not running the counterfactual experiment with unbiased context. However, tracing the model's internal computations can help us in distinguishing them.

Prospective Motivated Reasoning Detection The motivated and resistant cases both conflict with the hint in the unbiased context ($a_u \neq h$), and only differ in whether the model ultimately follows the hint ($a_h \stackrel{?}{=} h$). While in retrospect we can distinguish them by comparing their final answer, we want to study whether the model's internal representations can be used to predict whether the model will ultimately follow the hint, even before generating its CoT, to save computation and prevent the potentially unfaithful motivated reasoning.

2.3 Experimental Setup

Model. We conduct experiments with three language models spanning families of open-weight LLMs: 1) Qwen3-8B, 2) Llama-3.1-8B, and 3) Gemma-3-4B. We prompt the model to think step-by-step and then output its final choice letter. We sample with a temperature of 0 and set the maximum number of generation tokens to 2048.

Data. We use four multiple-choice question benchmarks: 1) MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021), 2) AQUA-RAT (Algebra Question Answering with Rationales) (Ling et al., 2017), 3) ARC (AI2 Reasoning Challenge) (Clark et al., 2018), and 4) CommonsenseQA (Talmor et al., 2019).

Hints. We use three different types of hints to construct biased contexts: 1) Sycophancy, 2) Consistency, and 3) Metadata. Table 1 explains the format in which the hints is provided, along with an example for each.

Hint	Description	Example
Sycophancy	Suggestion from an expert	"An expert indicates the answer is A."
Consistency	Model's prefilled response	User: <question> Assistant: The answer is (A). User: Explain your reasoning. Then give a final answer. Assistant:
Metadata	Embedded in XML metadata	<question >...</ question><question−metadata><correct−choice>A</correct−choice></question−metadata>

Table 1: We use three hints: sycophancy, consistency, and metadata.

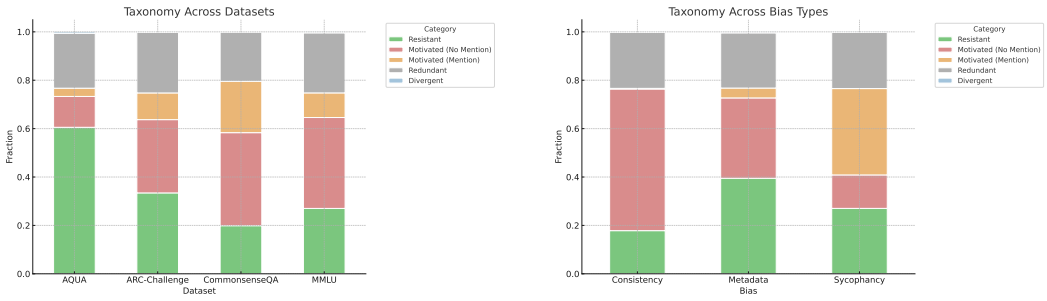

Figure 2: Transitions of model's answer due to the bias (left) across datasets, (right) across hints.

Probing. To analyze motivated reasoning, we extract residual stream activations after each attention block, at multiple generation steps of c_h. We train a Recursive Feature Machine (RFM) probe (Beaglehole et al., 2025) on these representations, with one probe per layer and step in the CoT. These kinds of non-linear probes are shown to be able to extract interpretable features in previous works. We use data extracted from %80 of the randomly selected questions for training and the rest as held-out.

3 Taxonomy of Responses to Bias

In this section, we examine the model's answer transitions from the unbiased context to the biased context. For every question in a dataset we construct an *unbiased prompt* x_u containing only the multiple-choice question, and a *biased prompt* x_h for each of the possible choices h. The model's responses (c_u, a_u) and (c_h, a_h) are then collected to categorize outcomes into the four transition types. Note that since every possible option is given as a hint, the total number of resits and motivated cases is always equal to $\#choices$-1 times the number of the coincident (redundant) and divergent cases.

Hint Mentioning We also check whether the hint is mentioned in the responses or not, and split the motivated cases into mention and no-mention subsets. To check mentioning, we just filter keywords such as 'expert', 'hint', 'metadata', etc. We also did ask 'gpt-5-nano' to label a subset of the data as either articulating influence of the hint or not. We noted that keyword filtering filters almost all the positive cases along with some others. Therefore, for our purpose which is to avoid false negatives, keyword filtering was sufficient.

3.1 Results

In Figure 2 you can see the taxonomy of the cases, with the motivated cases divided into the mention and no-mention cases. We note that for more objective datasets such as AQUA, which is composed of algebraic tasks, the rate of resistance is much higher than the more subjective datasets such as CommonsenseQA. Also, the consistency bias is the most effective in motivating the model, while the metadata is the least effective. In spite of that, the rate of mentioning for the metadata case is less than the sycophancy case.

4 Detecting Hint-Motivated Reasoning

In this section, we retain the cases that the hint was not mentioned, so the CoT cannot give information about the reliance of the model on the hint. Therefore, without running counterfactual experiments, studying the internal computations of the model would be the only viable approach.

Warmup: Hint Prediction We start with training a probe over the internal representations of the model at the end of CoT that predicts the hint. In this case we also remove the cases where the model is answering the hinted option. Therefore, there is no reference to the hint at the end of CoT.

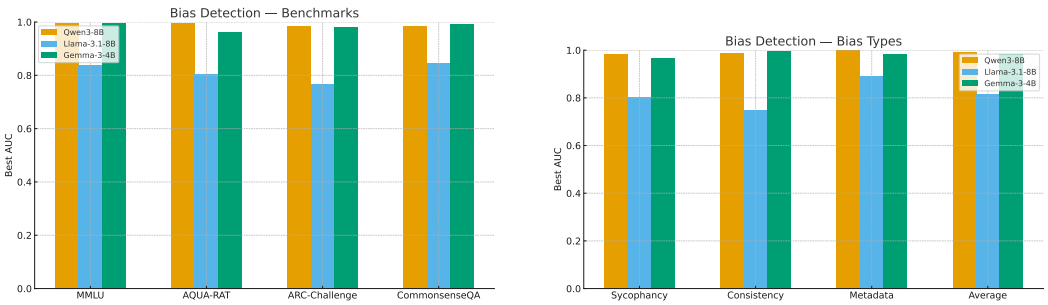

Figure 3: Best bias detection AUC across layers at the end of CoT (left) across datasets, (right) across hints.

Despite that, we note that the probe perfectly predicts the hint. This shows the possiblity of extracting information from internal representations of the model about its internal computations (See Figure 3).

Reliance on the hint is detectable at the end of CoT. The retrospective motivated reasoning 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 a reliable AUC at the end of CoT (See Figure 4). This shows the possibility of detecting the model's reliance on the hint, even though its CoT does not mention or follow the hint.

Switching to the hint is detectable before CoT generation. The prospective motivated reasoning detection probe that is trained to predict whether the model follows a hint that contradicts the model's unbiased answer, achieves a reliably high AUC before CoT generation (See Figure 5). This shows the possibility of detecting motivated reasoning intention from the internal representations of the model, even before generating CoT.

5 Related Work

Faithfulness of Chain-of-Thought Reasoning. A large body of work has shown that chain-of-thought (CoT) explanations produced by LLMs are not always faithful to the underlying computations. Early studies revealed that models may rationalize biased or hint-driven answers without acknowledging the true cause of their decision Turpin et al. (2023b); Lanham et al. (2023). Recent evaluations on reasoning-specific models confirm this gap: even when they rely on misleading cues, they rarely verbalize their influence Chen et al.; Chua & Evans (2025b); Arcuschin et al. (2025). Methods such as causal interventions on explanations Matton et al. (2025); Tutek et al. (2025) and mediation analyses Paul et al. (2024) quantify this unfaithfulness more precisely, while other work highlights the inherent hardness of eliciting faithful CoT from current models Tanneru et al. (2024).

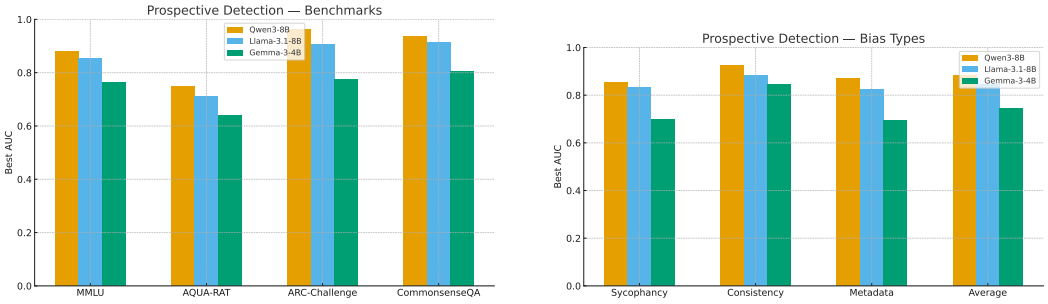

Figure 4: Best retrospective motivated reasoning Detection detection AUC across layers at the end of CoT (left) across datasets, (right) across hints.

Several approaches aim to increase the alignment between internal reasoning and verbalized CoT. These include debiasing strategies such as bias-augmented consistency training Chua et al. (2025), inference-time interventions like probabilistic dual-reward inference Li et al. (2025), and activation-level methods such as patching or control Yeo et al. (2024); Zhao et al. (2025). Frameworks like FRODO Paul et al. (2024) and FUR Tutek et al. (2025) provide structured ways to evaluate or improve reasoning faithfulness, while recent work investigates the limitations of fine-tuning, in-context learning, and activation editing Tanneru et al. (2024).

Mechanistic Interpretability and Probing. Another direction focuses on the internal representations of LLMs. Mechanistic interpretability efforts have mapped reasoning circuits and attribution graphs Lindsey† et al.; Sharkey et al. (2025). Latent knowledge studies aim to recover what models know but may not say Burns et al. (2024); Mallen et al. (2024), while probing methods test whether logical or causal structures can be extracted from representations Manigrasso et al. (2024); Cencerrado et al. (2025). Recent work identifies "thought anchors"—intermediate reasoning steps that disproportionately influence outcomes Bogdan et al. (2025).

Biases, Sycophancy, and Motivated Reasoning. Beyond faithfulness, models also exhibit cognitive biases similar to humans. Studies show that persona-assigned LLMs demonstrate motivated reasoning aligned with identity or ideology Dash et al. (2025), while others document sycophancy, where human preference training encourages models to echo user beliefs over truth Sharma et al. (2025).

6 Discussion and Limitations

In this paper, we focused on motivated reasoning as a 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. Moreover, the predictive features need to be investigated more deeply to understand the nature of internal computations that lead to motivated reasoning.

7 LLM Usage

We used Large Language Models for rewriting some paragraphs, and generate plots that are presented in the work. Moreover, in our experiments we used them to implement some methods that we then verified.

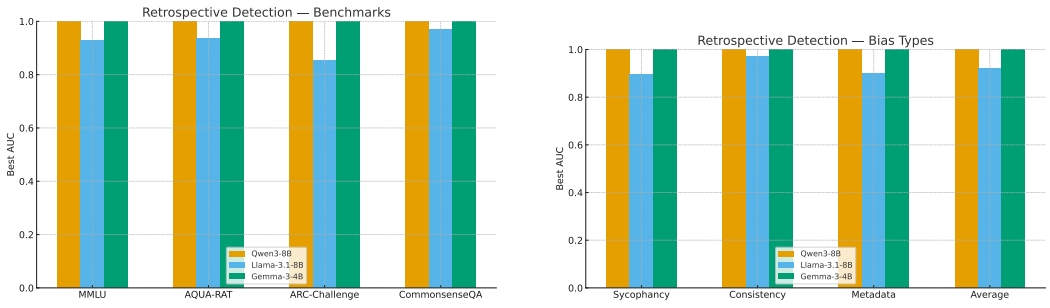

Figure 5: Best prospective motivated reasoning Detection detection AUC across layers before CoT generation (left) across datasets, (right) across hints.

REFERENCES

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful, June 2025. URL http://arxiv.org/abs/2503.08679. arXiv:2503.08679 [cs].

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.

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

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought Anchors: Which LLM Reasoning Steps Matter?, August 2025. URL http://arxiv.org/abs/2506.19143. arXiv:2506.19143 [cs].

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, March 2024. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827 [cs].

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No Answer Needed: Predicting LLM Answer Accuracy from Question-Only Linear Probes, September 2025. URL http://arxiv.org/abs/2509.10625. arXiv:2509.10625 [cs].

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don't Always Say What They Think.

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.

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

James Chua and Owain Evans. Are DeepSeek R1 And Other Reasoning Models More Faithful?, February 2025b. URL http://arxiv.org/abs/2501.08156. arXiv:2501.08156 [cs].

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought, June 2025. URL http://arxiv.org/abs/2403.05518. arXiv:2403.05518 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Saloni Dash, Amélie Reymond, Emma S. Spiro, and Aylin Caliskan. Persona-Assigned Large Language Models Exhibit Human-Like Motivated Reasoning, June 2025. URL http://arxiv.org/abs/2506.20020. arXiv:2506.20020 [cs].

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought, 2025. URL https://arxiv.org/abs/2410.08633.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning, July 2023. URL http://arxiv.org/abs/2307.13702. arXiv:2307.13702 [cs].

Jiazheng Li, Hanqi Yan, and Yulan He. Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6850–6866, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.340. URL https://aclanthology.org/2025.acl-long.340/.

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.

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. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Authors 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*‡ Affiliations Anthropic Published March 27. On the Biology of a Large Language Model. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation : Learning to solve and explain algebraic word problems, 2017. URL https://arxiv.org/abs/1705.04146.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting Latent Knowledge from Quirky Language Models, August 2024. URL http://arxiv.org/abs/2312.01037. arXiv:2312.01037 [cs].

Francesco Manigrasso, Stefan Schouten, Lia Morra, and Peter Bloem. Probing LLMs for Logical Reasoning. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner (eds.), *Neural-Symbolic Learning and Reasoning*, pp. 257–278, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71167-1. doi: 10.1007/978-3-031-71167-1_14.

Katie Matton, Robert Osazuwa Ness, John Guttag, and Emre Kıcıman. WALK THE TALK? MEASURING THE FAITHFULNESS OF LARGE LANGUAGE MODEL EXPLANATIONS. 2025.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh.

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.

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. URL https://arxiv.org/abs/2501.19393.

Franz Nowak, Anej Svete, Alexandra Butoi, and Ryan Cotterell. On the representational capacity of neural language models with chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2406.14197.

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. URL https://openreview.net/forum?id=iedYJm92o0a.

OpenAI. Learning to reason with llms, September 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning, October 2024. URL http://arxiv.org/abs/2402.13950. arXiv:2402.13950 [cs].

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open Problems in Mechanistic Interpretability, January 2025. URL http://arxiv.org/abs/2501.16496. arXiv:2501.16496 [cs].

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models, May 2025. URL http://arxiv.org/abs/2310.13548. arXiv:2310.13548 [cs].

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models, July 2024. URL http://arxiv.org/abs/2406.10625. arXiv:2406.10625 [cs].

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.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

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, 2023a.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, December 2023b. URL http://arxiv.org/abs/2305.04388. arXiv:2305.04388 [cs].

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps, June 2025. URL http://arxiv.org/abs/2502.14829. arXiv:2502.14829 [cs].

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.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. Towards Faithful Natural Language Explanations: A Study Using Activation Patching in Large Language Models, November 2024. URL http://arxiv.org/abs/2410.14155. arXiv:2410.14155 [cs].

Zekai Zhao, Qi Liu, Kun Zhou, Zihan Liu, Yifei Shao, Zhiting Hu, and Biwei Huang. Activation Control for Efficiently Eliciting Long Chain-of-thought Ability of Language Models, May 2025. URL http://arxiv.org/abs/2505.17697. arXiv:2505.17697 [cs].