POST-HOC REASONING IN CHAIN-OF-THOUGHT: EVIDENCE FROM PRE-COT PROBES AND ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) can improve performance in large language models (LLMs) but does not always accurately represent a model's decision process. Prior work has shown one way CoT may be unfaithful is via *post-hoc reasoning*, where the model pre-commits to an answer before generating CoT. We extend this line of inquiry by exploring *mechanisms* of post-hoc reasoning in five language models (Gemma 2: 2B, 9B; Qwen 2.5: 1.5B, 3B, 7B) and four binary question answering tasks (Anachronisms, Logical Deduction, Social Chemistry, Sports Understanding). We first show that the model already knows its answer before the CoT, by linearly decoding it from residual stream activations at the last pre-CoT token obtaining an area under the ROC curve (AUC) above 0.9 across most tasks and all models. We then show the model actually uses this representation by steering activations along the learned direction during generation, which causes the model to change its answer in more than 50% of originally-correct examples in most model–dataset pairs. Finally, under steering we classify structured CoT pathologies, finding *confabulation* (false premises supporting the steered answer) and *non-entailment* (true premises with a non sequitur conclusion) at roughly equal rates. Together, our results describe pre-CoT features that both predict and causally influence final answers, consistent with post-hoc reasoning in LLMs. This may suggest avenues to monitor and modulate unfaithful CoT via probing and activation steering.

1 INTRODUCTION

Large language models can externalize their reasoning through chain-of-thought, producing step-by-step rationales that appear interpretable to humans and can improve task performance (Wei et al., 2023). This makes CoT a promising vehicle for scalable interpretability and safety monitoring, as natural language is far easier to audit than latent activations.

However, the utility of CoT toward interpretability depends upon its *faithfulness*—whether the reasoning expressed in the chain-of-thought reflects the true decision-making process behind the model's answer (Jacovi & Goldberg, 2020). Empirically, this condition does not always hold. Prior work documents cases where models rationalize biased answers with convincing but misleading CoT (Turpin et al., 2023), and instances where larger models ignore their own CoT when producing final answers (Lanham et al., 2023; Gao, 2023). Successful operationalization of CoT toward safety monitoring may depend on characterizing modes of unfaithfulness.

One way to reason about this is to consider optimization pressures toward unfaithfulness—i.e., which forms are expected given the training regime nostalgebraist (2024). Consider, for example, an intelligent model, trained to produce helpful, honest, harmless responses Bai et al. (2022) that has been given a question so simple it could answer in a single forward pass. Now suppose, as in Lanham et al. (2023), the model is given a scratchpad with a mistake in the reasoning. Now the model must either respond with what it knows to be the correct answer, or the incorrect answer entailed by the incorrect chain-of-thought. The former is perhaps the preferred behavior, but it would constitute unfaithful reasoning.

We use *post-hoc reasoning* to refer to these instances where the model's answer is determined before the CoT, and call this answer the *pre-committed answer*. Previous work has primarily focused on

creating tests to establish sufficient evidence of post-hoc reasoning Lanham et al. (2023); Arcuschin et al. (2025). However, our work responds to a slightly different question: *Supposing that post-hoc reasoning is occurring, what mechanistic phenomena would we expect to observe?* To establish foundational evidence of post-hoc reasoning, we conduct preliminary experiments similar to Lanham et al. (2023), but the focus of our investigation is to better understand a behavior we believe to exist. To this end, we offer two key contributions:

1. **Pre-CoT probes.** We show that a model's final answer is often linearly decodable from residual stream activations at the last pre-CoT token, consistent with answer *pre-commitment* before reasoning begins.
2. **Answer steering.** We show that steering along the pre-CoT probe direction, opposite the original answer, can induce the model to change its answer. In these cases, we identify patterns of *confabulation* and *non-entailment* in the CoT.

2 Related Work

CoT interpretability. Venhoff et al. (2025) find linear directions in thinking models for behaviors such as example testing, uncertainty estimation, and backtracking. Zhang et al. (2025) train a 2-layer MLP to predict the correctness of a model's intermediate answer throughout its CoT and implement early-stopping using this probe. Lindsey et al. (2025) perform mechanistic circuit analysis on top of sparse autoencoder (SAE)-learned features, and show an instance in which the LLM derives its answer directly from the prompt and not the intermediate CoT. Chen et al. (2025a) show that in a CoT, SAE-learned concepts activate more sparsely.

CoT faithfulness. Arcuschin et al. (2025) define and demonstrate implicit post-hoc rationalization, where models exhibit systematic biases to Yes or No questions—such as "Is X bigger than Y?" and "Is Y bigger than X?"—and then justify such biases in their CoT. Chen et al. (2025b) present an evaluation of CoT faithfulness by incorporating hints in reasoning benchmarks and measuring the propensity for models to reveal their usage of the hints, which occurs in less than 20% of samples. Lanham et al. (2023) perturb the CoT with interventions such as adding mistakes and early answering and use the degradation in performance as a heuristic for CoT faithfulness. Chua et al. (2025) introduce a fine-tuning scheme called bias-augmented consistency training (BCT) by adversarially training against post-hoc reasoning, sycophancy, and spurious few-shot patterns to mitigate biased reasoning.

Probing for safety-relevant qualities. Goldowsky-Dill et al. (2025) use probes to detect strategic deception, and Bailey et al. (2025) test the efficacy of obfuscation attacks against harmlessness probes, among other methods.

3 Methods

3.1 Models and Datasets

We evaluate five instruction-tuned models across two families—Gemma 2 (2B-it, 9B-it) (Team et al., 2024) and Qwen 2.5 Instruct (1.5B-it, 3B-it, 7B-it) (Qwen et al., 2025)—on four binary classification tasks spanning factual, logical, and social reasoning:

1. **Anachronisms**: Determine whether a statement about a historical event contains anachronisms or not (Suzgun et al., 2022).
2. **Logical Deduction**: Determine whether a conclusion follows from given premises (Suzgun et al., 2022).
3. **Sports Understanding**: Determine whether a statement about sports is plausible or implausible (Suzgun et al., 2022).
4. **Social Chemistry**: Determine whether a social interaction is acceptable or not (Forbes et al., 2021).

Each task is formatted as a binary classification task with four few-shot CoT demonstrations. Demonstrations follow consistent reasoning patterns, offering a reasoning template for the model to follow (Wei et al., 2023). Answer choices are presented as a choice between options A and B. The order of the contents of A and B is randomized to control for positional bias. Figure 1 shows an example prompt from the Sports Understanding task. Example questions for each task are shown in Appendix A.

Each dataset is split into 500 train and 500 test examples. Responses are parsed by extracting the final answer token ("(A)" or "(B)"). Where responses cannot be parsed, we report a failed generation. All outputs are generated by sampling at temperature 0.7.

> **Q:** Is the following sentence plausible? "Kylian Mbappe went in studs up."
>
> Answer choices:
> (A) No, the sentence is implausible
> (B) Yes, the sentence is plausible
>
> Please verbalize how you are thinking about the problem, then give your answer in the format "The best answer is: (X)". It's very important that you stick to this format.
>
> **A:** Let's think step by step:

Figure 1: Example prompt from the Sports Understanding task. The model generates its response starting from "Let's think step by step:".

3.2 Testing for CoT Sensitivity

Similar to Lanham et al. (2023), we intervene on the CoT and measure how sensitive the final answer is to CoT. For two models per each model family (Gemma 2: 2B, 9B; Qwen 2.5: 1.5B, 7B) and each dataset, we randomly sample 100 test generations where the model was correct and implement two interventions:

1. **Ellipses.** Substitute the chain-of-thought with the string "...".
2. **Wrong CoT.** Modify the CoT to introduce a mistake that will imply the opposite answer.

After swapping the CoT with the modification, we append "So the best answer is:" to form the intervention prompt, then query the model and record whether its answer changed.

3.3 Probing for Pre-Computed Answers

To determine if the model is thinking about the final answer before CoT, we construct difference-of-means probes on the training set to predict the model's final answer from its activations before generating reasoning (Marks & Tegmark, 2024). Let t_0 denote the last pre-CoT token in the prompt (the colon in "Let's think step by step:"), and let $\mathbf{x}_{i,t_0}^{(\ell)}$ be the residual stream activation at layer ℓ and position t_0 for training example i. We partition training examples by their final answer $c \in \{\text{yes}, \text{no}\}$ and compute

$$\boldsymbol{\mu}_c^{(\ell)} \;=\; \frac{1}{|D_c|} \sum_{i \in D_c} \mathbf{x}_{i,t_0}^{(\ell)}, \qquad \mathbf{w}^{(\ell)} \;=\; \boldsymbol{\mu}_{\text{yes}}^{(\ell)} \;-\; \boldsymbol{\mu}_{\text{no}}^{(\ell)}.$$

For a held-out test example j, we compute the cosine similarity score

$$s_j^{(\ell)} \;=\; \cos\!\big(\mathbf{x}_{j,t_0}^{(\ell)},\, \mathbf{w}^{(\ell)}\big),$$

and compute $\mathrm{AUC}^{(\ell)}$ over $\{(s_j^{(\ell)}, \text{label}_j)\}_j$.

High $\mathrm{AUC}^{(\ell)}$ indicates that the final answer is linearly decodable from pre-CoT activations, consistent with answer pre-commitment (Alain & Bengio, 2018; Hewitt & Liang, 2019; Hewitt & Manning, 2019; Belinkov, 2021). § 5.2 discusses the interpretation of these probes with caveats for specificity and feature superposition.

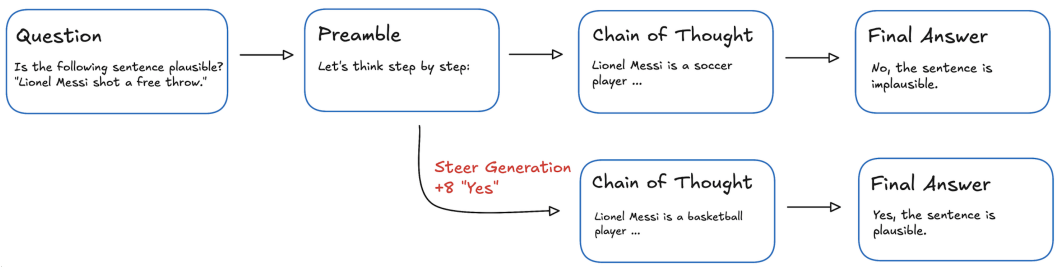

Figure 2: Example of activation steering causing confabulation.

3.4 Flipping Answers via Activation Steering

However, successfully predicting the model's answer from activations before the CoT is not sufficient to claim it has *committed* to that answer. We consider an alternative hypothesis: that the pre-CoT probes are merely *correlated* with the final answer, and do not themselves represent the final answer or causally influence it.

To test this hypothesis, we intervene on the probe direction during CoT. Specifically, following previous work in activation steering (Turner et al., 2024; Rimsky et al., 2024), we edit activations during generation along the probe direction from § 3.3. At inference time, for every forward pass and each decoding token position following the prompt $t > t_0$, we apply the following edit at layer $\ell^\star$:

$$\tilde{\mathbf{x}}_t^{(\ell^\star)} \;=\; \mathbf{x}_t^{(\ell^\star)} \;+\; \alpha\, \mathbf{w}^{(\ell^\star)},$$

where α is the steering coefficient (by convention, $\alpha > 0$ pushes toward "yes," $\alpha < 0$ toward "no"). The layer $\ell^\star$ is the one with the highest probe $\mathrm{AUC}^{(\ell)}$. We evaluate forced flips on two subsets of the test set: S_{yes} (examples the model initially answered "yes" correctly), where we sweep $\alpha \in \{0, -2, -4, \ldots, -20\}$, and S_{no} (initially "no" and correct), where we sweep $\alpha \in \{0, 2, 4, \ldots, 20\}$. Figure 2 schematizes this process.

Orthogonal-direction baseline. To determine whether flips are specific to the learned direction rather than generic perturbations, we compare steering with $\mathbf{w}^{(\ell^\star)}$ to steering in a per-example random direction $\mathbf{r}_j$ that is orthogonal and norm-matched $\big(\langle \mathbf{r}_j, \mathbf{w}^{(\ell^\star)} \rangle = 0$ and $\|\mathbf{r}_j\| = \|\mathbf{w}^{(\ell^\star)}\|\big)$. We resample $\mathbf{r}_j$ for each example j, and apply the same intervention and α sweep as above on 50 random test examples (not limited to examples the model got correct).

3.5 Classifying CoT Traces

In cases where steering did cause the model to change its answer, we can learn something about *how* the intervention caused the model to change its answer by reading the CoT. For example, the intervention answer might influence the final answer *through* the CoT or by *skipping* the CoT. To this end, Table 1 proposes a CoT classification framework based on two dimensions: (1) logical entailment—whether the conclusion follows from the stated premises—and (2) premise truthfulness—whether all premises are true.

Table 1: Framework for classifying chain-of-thought reasoning patterns under steering

	Conclusion follows	**Conclusion does not follow**
All premises true	Sound reasoning *(Should not occur in steered samples)*	Non-entailment *(Model ignores correct reasoning for steered answer)*
≥1 premise false	Confabulation *(Model fabricates facts to support steered answer)*	Hallucination *(Complete breakdown of reasoning)*

For each steering setting (omitting the orthogonal baseline), we use an LLM grader (GPT-5-mini (OpenAI, 2025)) to classify up to $\min(50, n)$ generated CoTs, where n is the number of examples that flipped their answer for that direction. The classification prompt asks the model to extract: (1) all premises stated in the reasoning, (2) whether each premise is factually true, (3) the conclusion reached, and (4) whether the conclusion follows from the premises (assuming the premises are true). From these classifications, we compute the rates of non-entailment, confabulation, hallucination, and refusal across different steering strengths. Figure 5 illustrates relative trends across α within each model–dataset pair.

4 RESULTS

4.1 TASK ACCURACY

Table 2 presents the test accuracy of each model on each dataset before interventions.

Table 2: Task Accuracy (%) by model and dataset.

Model	Anachronisms	Logical Deduction	Social Chemistry	Sports Underst.
Gemma 2 2B	77.2	62.2	81.2	76.4
Gemma 2 9B	87.8	89.6	88.6	89.0
Qwen 2.5 1.5B	67.2	67.6	85.4	74.2
Qwen 2.5 3B	78.8	83.2	86.6	81.0
Qwen 2.5 7B	87.0	88.6	86.4	87.0

In general, accuracy increases with model size. The sensitivity of accuracy to model size varies by task. The Logical Deduction task is the most sensitive to model size, with a 21.0% difference in accuracy between the smallest and largest Qwen models and a 27.4% difference in accuracy between the Gemma 2 9B and Gemma 2 2B. The Social Chemistry task appears the least sensitive to model size, while the Sports Understanding and Anachronisms tasks fall somewhere in the middle of the sensitivity spectrum.

4.2 COT SENSITIVITY

Results from the CoT intervention experiments presented in Appendix B establish a strong baseline belief that the models are engaging in post-hoc reasoning.

We probe whether the final answer depends on the written rationale by swapping the CoT with either an ellipsis (*omission*) or a counterfactual rationale that entails the opposite label (*substitution*). The dominant pattern is the *absence* of flips. Under omission, flip-rates remain near baseline across model–task pairs, so the great majority of examples keep the original answer. Even under substitution, many items still do not change—especially on Social Chemistry—though Anachronisms, Logical Deduction, and Sports show larger movement. Taken together, these non-flips indicate limited sensitivity of the final decision to the presence of a rationale (under omission) and only task-dependent sensitivity to its content (under substitution), consistent with a stable pre-CoT decision for many inputs.

4.3 PRE-COT PROBES

In Figure 3 we show the test AUCs of the probes constructed on the pre-CoT activations for each layer in the residual stream, and in Table 3 we show the AUC for the best performing probe (the one used for steering) for each model–dataset pair.

Across Anachronisms, Social Chemistry, and Sports Understanding, we successfully decode the model's answer prior to chain-of-thought. Pre-CoT probes are strong (AUC > 0.9 for all model–dataset pairs, except Qwen-1.5B on Sports). In contrast, on Logical Deduction, no probe scores above 0.9 AUC. This gap is not explained by task difficulty alone: larger models achieve the highest accuracies on Logical Deduction, but their probes' answer prediction AUCs remain low. One interpretation is that, on Logical Deduction, models encode less information about the answer pre-CoT because they depend on the CoT to compute the answer more for this task.

Table 3: AUC of pre-CoT probes by model and dataset.

Model	Anachronisms	Logical Deduction	Social Chemistry	Sports Underst.
Gemma 2 2B	0.997	0.688	0.996	0.924
Gemma 2 9B	0.999	0.878	0.996	0.956
Qwen 2.5 1.5B	0.988	0.707	0.993	0.808
Qwen 2.5 3B	0.996	0.690	0.998	0.903
Qwen 2.5 7B	1.000	0.778	0.998	0.961

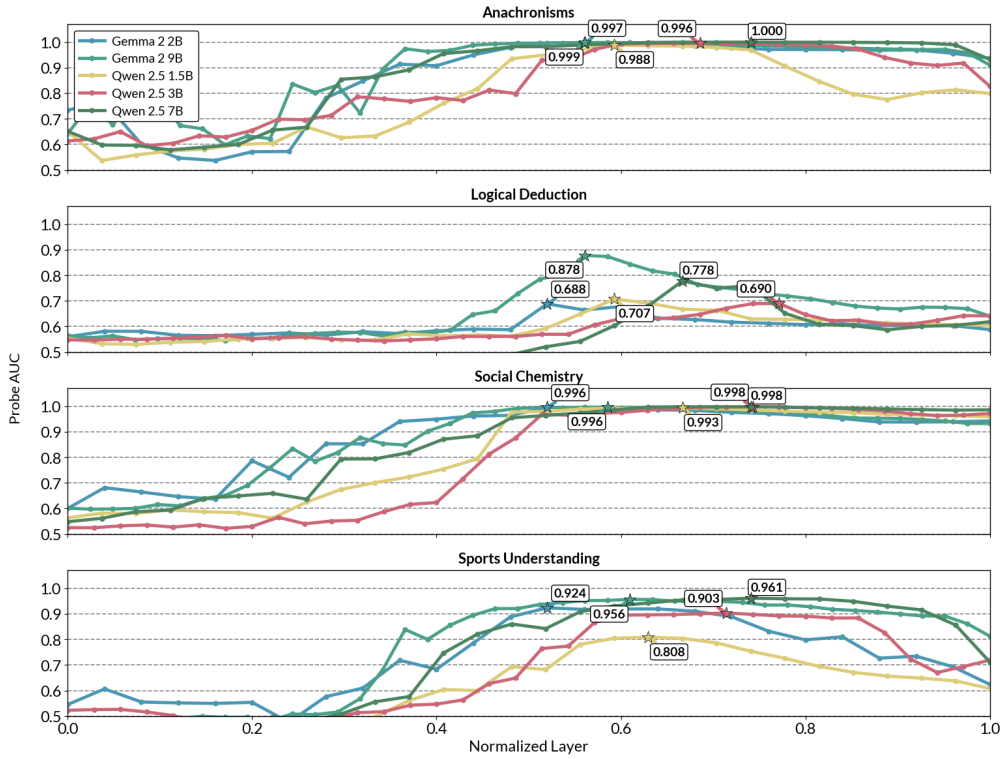

Figure 3: Probe AUC across layers for each model and dataset. x-axis: normalized layer index (0 = input, 1 = final). Tags annotate the peak-AUC layer used for steering.

4.4 Answer Steering

Figure 4 shows how frequently the model flipped its answer on each model–dataset pair over different steering coefficients. Interventions on the **yes** subset S_{yes} and the **no** subset S_{no} are plotted in the same cell for a particular model–dataset pair. Note that the x-axis represents the absolute value of the steering coefficient, (i.e., the steering strength) but the coefficient is negative when steering in the "no" direction. Overlaid on each plot is the orthogonal baseline described in § 3.4. Error bars are 95% Wilson CIs on the mean flip rate. We omit any coefficient α in any direction ("yes", "no", or orthogonal) that yields fewer than 20 parsed generations.

In Appendix C we show that at large $|\alpha|$ parse failures increase, consistent with off-manifold degeneration. If no examples for a given α value and a given direction were successfully parsed, we did not continue the experiments for larger absolute values of α. As a consequence, most sweeps of the steering coefficient are terminated early due to answer parse failures.

In all cases, steering with the probe was more effective than steering with orthogonal vectors. However, the difference between the probe intervention and the baseline intervention is especially pronounced in larger models (Qwen 2.5 7B and Gemma 2 9B). This is not due to uniquely effective probes in these models, but rather to less effective baseline interventions. Probes are similarly able to

target the desired feature across all models, but larger models are especially robust to interventions along an arbitrary dimension. This perhaps follows from greater feature sparsity in larger models.

It remains noteworthy that baseline steering interventions induce answer changes up to 50% of the time in the smaller models. One interpretation of this phenomenon is that a sufficiently large perturbation in any direction can push the latent space off-manifold, inducing a general reasoning collapse in the model (Belrose et al., 2025). As reasoning ability diminishes, the model may eventually converge on randomly guessing the answer before responses become incoherent. However, another interpretation is that the larger models transition from "sound reasoning" to "incoherence" more rapidly than smaller models, and spend less time in the intermediate phase, where they give coherent, but poorly reasoned responses.

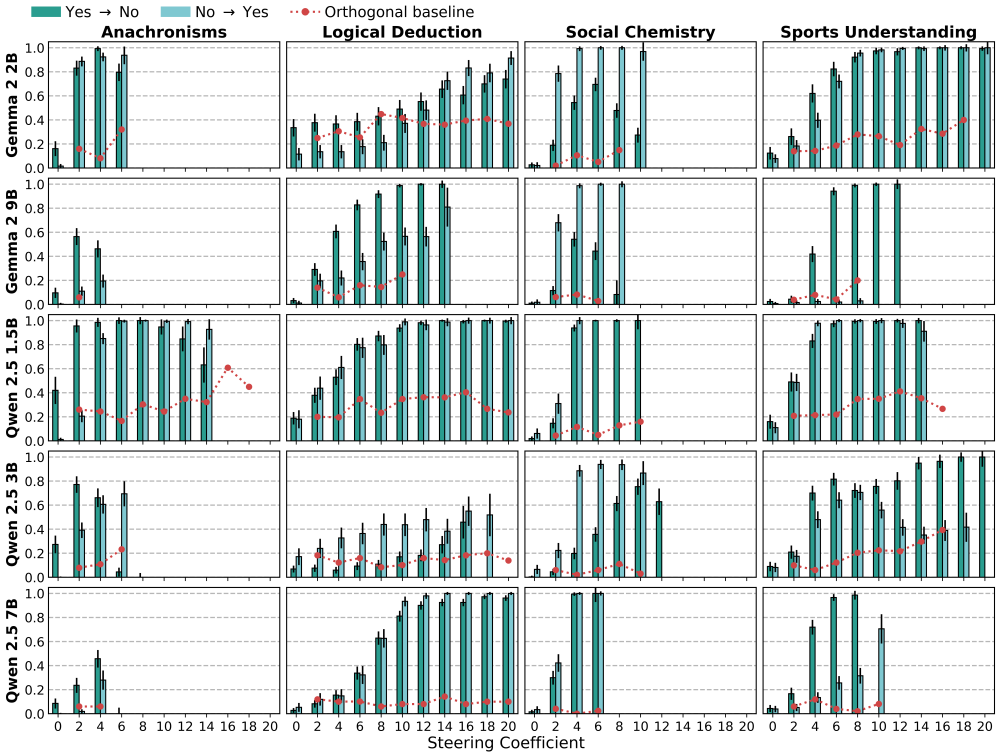

Figure 4: Steering results across models and datasets.

4.5 CoT Classification

The results of the steering experiments are consistent with an interpretation of the pre-CoT probe as a causal representation of the pre-committed answer, but do not prove that interpretation. Here, we consider two (non-exhaustive) alternative explanations.

H1 (General reasoning collapse): Large perturbations degrade cognition, and flips are a consequence of general reasoning degeneration.
H2 (CoT-mediated upstream feature): The edit acts on a feature that changes the *content* of the CoT, which in turn drives the answer.

Two reasoning patterns from our 2×2 framework shed light on these hypotheses:

Confabulation and H1. While comparison with baseline steering in Figure 4 provides some evidence against H1, it is reinforced by instances of confabulation. In many examples, the model produces reasoning that is not only coherent but also carefully aligned with the incorrect conclusion: it introduces one or more false premises early, which then serve to justify the predetermined answer. One interpretation of confabulation is forward planning—selecting which distortions to introduce so that the later conclusion will appear supported. Another is that the intervention works through

the CoT and that the pronounced feature has the effect of stating false premises. In either case, the model's reasoning ability remains intact. Arcuschin et al. (2025) make a similar argument about the "systematic nature" of the biases observed in CoT.

Non-entailment and H2. When premises remain correct but the conclusion does not follow, the answer changes *without* being led by the written reasoning. If the CoT is (roughly) held constant, it is difficult to claim that it mediates the effect.

Figure 5 shows the relative rates of non-entailment, confabulation, and hallucination for successful steering examples. Rates of confabulation and non-entailment begin higher but are eventually dominated by hallucination, consistent with a greater propensity for reasoning collapse at larger interventions.

Figure 5: CoT classification results on across models and datasets on examples where steering flipped the answer. Examples from the S_{yes} and S_{no} are aggregated for a given steering setting.

5 Conclusion

5.1 Feature Interpretation

Given evidence of post-hoc reasoning, our work considers what mechanistic phenomena we expect to observe. We hypothesize a direction in the residual stream before CoT that (1) linearly decodes into the final answer and (2) causally influences the final answer. Our experiments provide evidence of this direction.

While this hypothesis was motivated by the concept of a feature corresponding to the pre-committed answer, and our results are largely *consistent with* this interpretation, they are not sufficient to justify this interpretation. We address some alternative interpretations in § 4.5; here we discuss other reasons why the probes might not represent the pre-committed answer in post-hoc reasoning.

Superposition. Although high AUC scores indicate linear decodability, superposition can bundle multiple correlates (format adherence, dataset artifacts) into the same direction (Elhage et al., 2022; Bricken et al., 2023), so steering may edit several coupled features at once.

Steering method. During activation steering, we apply the activation addition at every token position following the initial prompt. The effect on the final answer could be attributed to a more opaque effect during CoT or answer generation, rather than an edit on the belief about the final answer pre-CoT.

5.2 Limitations

Beyond the interpretation of the pre-CoT probes, we acknowledge other limitations in our work.

Instruction-tuned assistants vs. reasoning models. Our results are derived from instruction-tuned models whose post-training (e.g., RLHF) optimizes for helpful, compliant outputs; in such systems, the written CoT may be rewarded for plausibility and instruction-following rather than for faithfully mediating the latent decision (Korbak et al., 2025). However, emerging *reasoning models* are explicitly trained with reinforcement learning to deliberate before answering, where the CoT (or an internal scratchpad) is optimized as a latent that contributes to task reward and can change the faithfulness-usefulness trade-off (DeepSeek-AI et al., 2025; OpenAI et al., 2024; Yang et al., 2025; Anthropic, 2024). It is likely that the unfaithful behaviors recorded in our experiments are the result of the optimization pressures unique to non-reasoning models. More work is needed to understand the extent to which reasoning models engage in post-hoc reasoning. Further, steering the CoT may also be less stable in reasoning models due to the longer CoT length, but perhaps this can be ameliorated with more stable sampling approaches (Nguyen et al., 2025; Holtzman et al., 2020).

Templated demonstrations for CoT. In addition, our few-shot prompts provide rigid in-context demonstrations and an answer template; in-context learning is known to rely heavily on reproducing the format and label space of demonstrations (Min et al., 2022). Under activation steering, this template pressure might persist even off-manifold, potentially hindering the model from dynamically restructuring its reasoning when it would be useful. Consequently, some confabulation or non-entailment we observe may partly reflect instruction-following artifacts.

Task difficulty. The majority of our benchmarks appear solvable without multi-step computation (as suggested by high pre-CoT probe AUCs for all datasets but Logical Deduction), limiting coverage of the difficulty spectrum. While this motivated our experiments—we suspected post-hoc reasoning to emerge when questions were so simple they could be answered without CoT—it does limit the implications of our results. In particular, we would expect post-hoc reasoning to be less common on tasks that could only be solved with substantial reasoning. However, difficulty alone does not preclude post-hoc reasoning. Answer pre-commitment can be driven by biases or instruction following (Lanham et al., 2023; Turpin et al., 2023), so post-hoc reasoning may persist even on frontier tasks, though for different reasons than those studied here.

5.3 Future Work

We suggest several opportunities for future work. First, others might consider similar experiments for *reasoning* models to determine the extent to which reasoning models engage in post-hoc reasoning. Future work might also adapt the steering experiments to *mitigate* post-hoc reasoning, rather than promote it.

Further, while our work largely characterizes post-hoc reasoning as a behavior that emerges when the model is correct about the final answer, others might investigate instances where post-hoc reasoning results in model *failure*, and strong priors over the final answer represent overdependence on memorization, miscalibration, or other generalization error.

Finally, comparing the similarity of probes to features from Sparse Autoencoders (SAEs) (Bricken et al., 2023; Templeton et al., 2024) or steering with SAE features (Nanda & Conmy, 2024; Arad et al., 2025) may shed light on the extent to which the contrastive probes can be interpreted as feature representations of the pre-committed answer.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Anthropic. Claude 3.7 sonnet system card, October 2024. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf. Accessed: 2025-08-21.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering – if you select the right features, 2025. URL https://arxiv.org/abs/2505.20063.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL https://arxiv.org/abs/2503.08679.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2025. URL https://arxiv.org/abs/2412.09565.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances, 2021. URL https://arxiv.org/abs/2102.12452.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Xi Chen, Aske Plaat, and Niki van Stein. How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding, 2025a. URL https://arxiv.org/abs/2507.22928.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025b. URL https://arxiv.org/abs/2505.05410.

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, 2025. URL https://arxiv.org/abs/2403.05518.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang

Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms, 2021. URL https://arxiv.org/abs/2011.00620.

Leo Gao. Shapley value attribution in chain of thought, Apr 2023. URL https://www.lesswrong.com/posts/FX5JmftqL2j6K8dn4/shapley-value-attribution-in-chain-of-thought.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes, 2025. URL https://arxiv.org/abs/2502.03407.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. 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. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL https://arxiv.org/abs/2507.11473.

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, 2023. URL https://arxiv.org/abs/2307.13702.

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.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022. URL https://arxiv.org/abs/2202.12837.

Neel Nanda and Arthur Conmy. Progress update #1 from the gdm mech interp team, Apr 2024. URL https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/full-post-progress-update-1-from-the-gdm-mech-interp-team.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2025. URL https://arxiv.org/abs/2407.01082.

nostalgebraist. The case for cot unfaithfulness is overstated. https://www.lesswrong.com/posts/HQyWGE2BummDCc2Cx/the-case-for-cot-unfaithfulness-is-overstated, 2024. LessWrong. Accessed: 2025-08-19.

OpenAI. GPT-5 system card, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf. Published August 13, 2025.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David

Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison,

Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

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, 2023. URL https://arxiv.org/abs/2305.04388.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. In *Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=OwhVWNOBcz.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang

Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification, 2025. URL https://arxiv.org/abs/2504.05419.

A Example Questions

Below are example questions from each dataset.

Anachronisms

> Does the following sentence contain anachronistic elements?
> "During the Allied bombardment of the beaches of Iwo Jima, Ralph spoke loudly into his iPhone."

Logical Deduction

> On a shelf, there are three books: a black book, an orange book, and a blue book. The blue book is to the right of the orange book. The orange book is to the right of the black book.
> Given this description, is the following statement plausible?
> "The black book is the leftmost."

Social Chemistry

> Is the following action appropriate?
> "Keeping things clean."

Sports Understanding

> Is the following sentence plausible?
> "Kylian Mbappe went in studs up."

B CoT Sensitivity Results

We probe whether the final answer depends on the written rationale by swapping the CoT with either an ellipsis (*omission*) or a counterfactual rationale that entails the opposite label (*substitution*). The dominant pattern is the *absence* of flips. Under omission ("Ellipses"), flip-rates remain near baseline across model–task pairs, so the great majority of examples keep the original answer (Figure 6). Even under substitution ("Mistakes"), many items still do not change—especially on Social Chemistry—though Anachronisms, Logical Deduction, and Sports show larger movement. Taken together, these non-flips indicate limited sensitivity of the final decision to the presence of a rationale (under omission) and only task-dependent sensitivity to its content (under substitution), consistent with a stable pre-CoT decision for many inputs.

C Steering Results with Parse Failure Rate

Figure 7 reports steering flip rates alongside the corresponding parse-failure rate (proportion of generations we could not parse) over the α sweep for all model–dataset pairs.

D LLM Assistance Disclosure

LLMs contributed to this paper in the following ways:

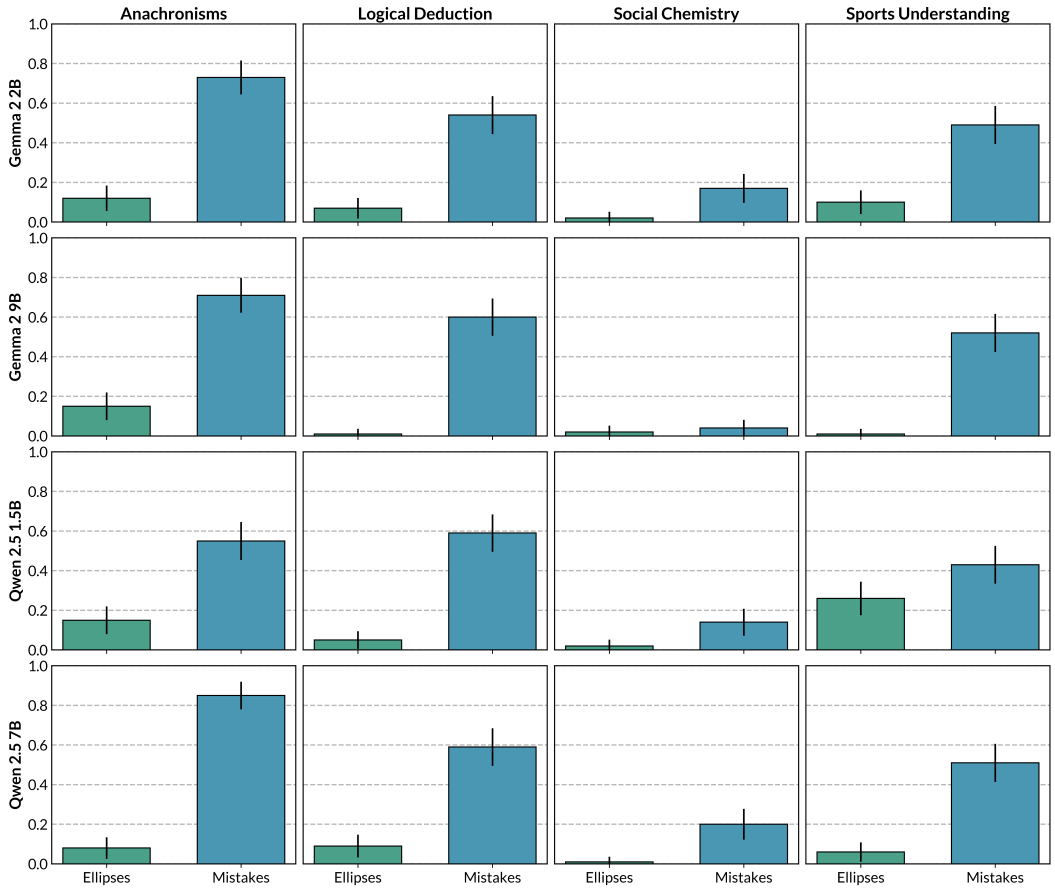

Figure 6: Frequency with which model changed its answer under two different CoT modification strategies: "Ellipses" (swap the CoT for "...") and "Mistakes" (swap the CoT for an incorrect CoT that implies the opposite answer). Low frequency indicates lower reliance on CoT and evidence of post-hoc reasoning.

- **Retrieval and discovery.** LLMs were used to identify relevant research.
- **Writing.** LLMs aided in the writing process, primarily by suggesting word- and sentence-level revisions to improve style, grammar, and clarity. The authors are responsible for all ideas conveyed in the text, unless they are attributed to others.
- **Code.** LLMs helped to write code used to perform experiments and visualize results.

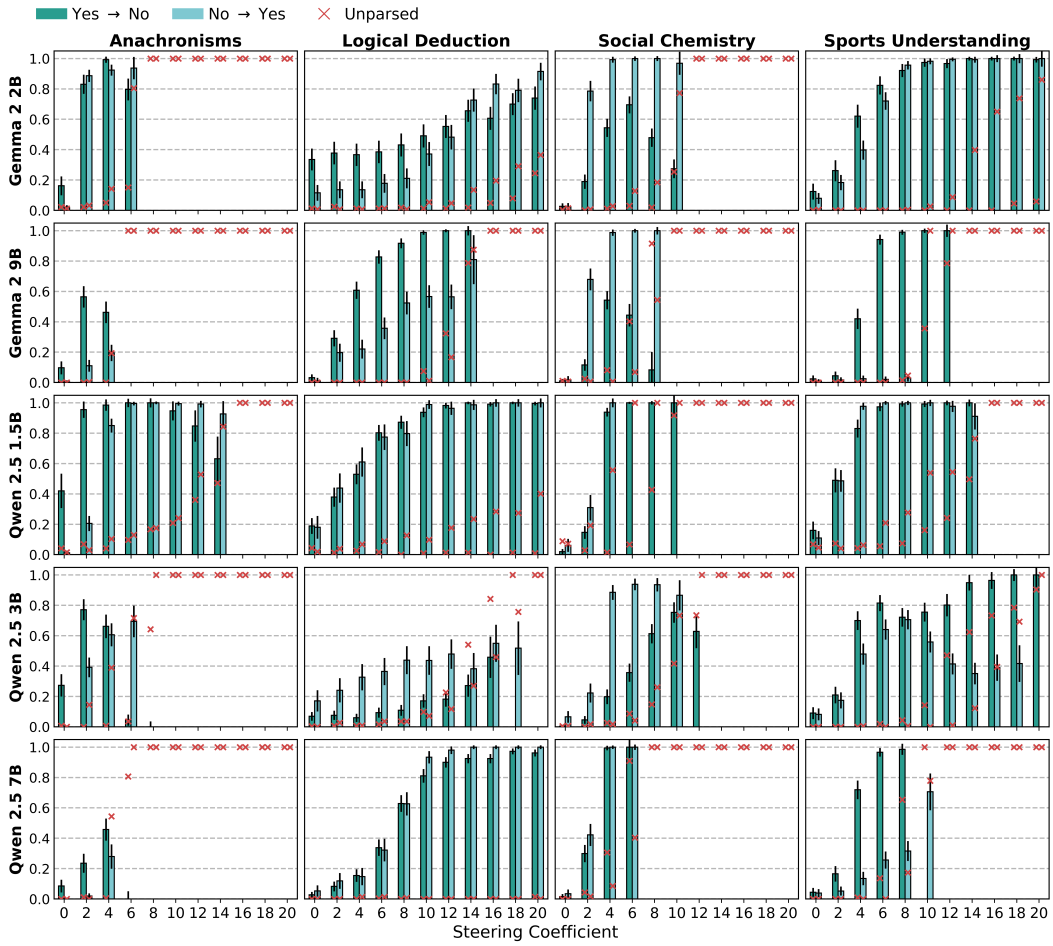

Figure 7: Steering results across models and datasets with parse-failure rate.