

# 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 for 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*.

Prior work has established evidence of post-hoc reasoning through primarily prompt-level experiments (Lanham et al., 2023; Arcuschin et al., 2025; Bao et al., 2024). For example, models might respond in the same way when their CoT is swapped with an incorrect CoT. These findings invite hypotheses about what *mechanistic phenomena* are involved in post-hoc reasoning.

Our experiments are sequenced in the following way.

**Empirical premise (P0).** Prior work has shown that on some reasoning tasks, models may "know" the answer prior to CoT and perform reasoning post-hoc. For example, models may respond correctly when CoT is removed, or replaced with a misleading CoT. We select datasets where CoT is differentially useful, and verify that our models exhibit this behavior on some tasks. In § 4.1 we compare the accuracy of our models with and without CoT, and in § 4.2 we evaluate model accuracy under two CoT interventions: removal (swapping with ellipses) and substitution (swapping with an incorrect, misleading CoT).

**Hypotheses.** Conditional on this premise, we test three hypotheses:

- **Representational pre-commitment (H1).** The model’s final answer is encoded in pre-CoT activations in the residual stream, and is linearly decodable by a simple probe (§ 4.3).
- **Causal pre-commitment feature (H2).** The probe direction is not merely predictive but causal: steering activations along this direction shifts the model’s answer far more than equally large orthogonal perturbations (§ 4.4).
- **Pathologies of unfaithfulness (H3).** When steered in the direction of the incorrect response, the model’s verbalized reasoning will exhibit two patterns: (1) stating false premises to support the steered answer (*confabulation*) and (2) stating true premises but giving a conclusion that does not follow (*non-entailment*) (§ 4.5).

## 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. Bao et al. (2024) use prompt interventions to construct causal models of CoT reasoning, identifying instances where models are "explaining" rather than reasoning about the answer. 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.

**Our contribution.** Prior work on CoT faithfulness has largely operated at the prompt level—e.g., modifying or swapping rationales to measure robustness of the final answer. We extend this work by studying post-hoc reasoning at the level of internal activations, probing and steering pre-CoT residual-stream features that predict and influence the final answer.

### 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 and offer 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

We evaluate how sensitive the model is to chain-of-thought in two ways:

**Accuracy improvement due to CoT.** We evaluate model accuracy with and without CoT. In the no-CoT examples, the model is instructed to respond only with the answer, including no reasoning. The in-context demonstrations for the no-CoT evaluation are the same as those for the CoT evaluation, but stripped of the CoT.

**CoT intervention.** 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 model family (Gemma 2: 2B, 9B; Qwen 2.5: 1.5B, 7B) and each dataset, we randomly sample 50 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.

The details of the intervention procedure are described in Appendix B.

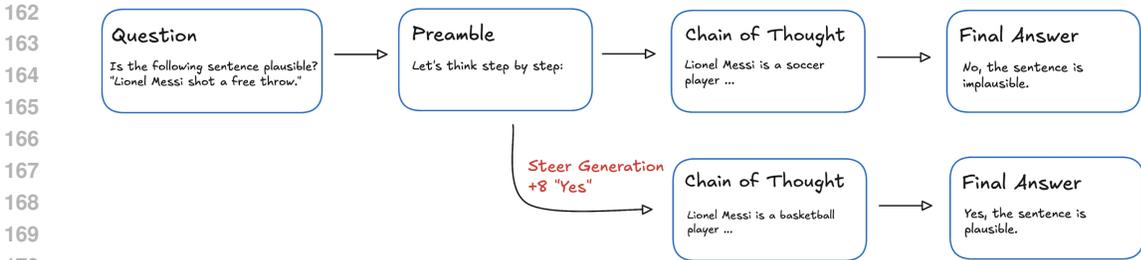


Figure 2: Example of activation steering causing confabulation.

### 3.3 PROBING FOR PRE-COMPUTED ANSWERS

To determine if the final answer is linearly decodable pre-CoT (**H1**), 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  $\ell$  and position  $t_0$  for training example  $i$ . We partition training examples by their final answer  $c \in \{\text{yes}, \text{no}\}$  and compute

$$\mu_c^{(\ell)} = \frac{1}{|D_c|} \sum_{i \in D_c} \mathbf{x}_{i,t_0}^{(\ell)}, \quad \mathbf{w}^{(\ell)} = \mu_{\text{yes}}^{(\ell)} - \mu_{\text{no}}^{(\ell)}.$$

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

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

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

where high  $\text{AUC}^{(\ell)}$  indicates that the final answer is linearly decodable from pre-CoT activations (Alain & Bengio, 2018; Hewitt & Liang, 2019; Hewitt & Manning, 2019; Belinkov, 2021).

### 3.4 FLIPPING ANSWERS VIA ACTIVATION STEERING

Supposing the directions identified in § 3.3 are predictive of the final answer, their interpretation is ambiguous. Specifically, it is unclear whether they are merely predictive of the final answer, or causally influence it (**H2**).

To test this hypothesis, we intervene on the probe direction during CoT. 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  $\ell^*$ :

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

where  $\alpha$  is the steering coefficient (by convention,  $\alpha > 0$  pushes toward “yes,”  $\alpha < 0$  toward “no”). The layer  $\ell^*$  is the one with the highest probe  $\text{AUC}^{(\ell)}$ . We evaluate forced flips on two subsets of the test set:  $S_{\text{yes}}$  (examples the model initially answered “yes” correctly), where we sweep  $\alpha \in \{0, -2, -4, \dots, -20\}$ , and  $S_{\text{no}}$  (initially “no” and correct), where we sweep  $\alpha \in \{0, 2, 4, \dots, 20\}$ . 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  $\mathbf{w}^{(\ell^*)}$  to steering in a per-example random direction  $\mathbf{r}_j$  that is orthogonal and norm-matched ( $\langle \mathbf{r}_j, \mathbf{w}^{(\ell^*)} \rangle = 0$  and  $\|\mathbf{r}_j\| = \|\mathbf{w}^{(\ell^*)}\|$ ). We resample  $\mathbf{r}_j$  for each example  $j$ , and apply the same intervention and  $\alpha$  sweep as above on 50 random test examples (not limited to examples the model got correct).

### 3.5 CLASSIFYING CoT TRACES

In instances where steering caused the model to change its answer, we hypothesize that the model’s verbalized reasoning will follow two patterns: (1) by inventing premises supporting the incorrect

answer (*confabulation*) and (2) ignoring a correct CoT to conclude the incorrect answer (*non-entailment*) (**H3**). In Table 1 we generalize this in a 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 ( <i>Should not occur in steered samples</i> )	Non-entailment ( <i>Model ignores correct reasoning for steered answer</i> )
$\geq 1$ premise false	Confabulation ( <i>Model fabricates facts to support steered answer</i> )	Hallucination ( <i>Complete breakdown of reasoning</i> )

We use GPT-5-mini (OpenAI, 2025) as an LLM grader to classify the reasoning traces of generations from § 3.4 where steering caused the model to respond with the incorrect answer. For each steering setting (combination of model, dataset, and steering coefficient  $\alpha$ ) we sample  $\min(50, n)$  generations for classification, where  $n$  is the number of examples that flipped their answer for that direction. We exclude steering settings where there are fewer than 20 examples to classify.

The classification prompt instructs the model to return two fields: (1) a boolean indicating whether the reasoning trace contains any false premises and (2) a boolean indicating whether the model’s final answer logically follows from the stated premises, assuming they are true. Classifications are computed from these two fields according to the schema in Table 1. More details about the classification prompt are given in Appendix F.

## 4 RESULTS

### 4.1 TASK ACCURACY

Table 2 presents the test accuracy of each model on each dataset with and without chain-of-thought.

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

Model	Anachronisms		Logical Deduction		Social Chemistry		Sports Underst.	
	No CoT	CoT	No CoT	CoT	No CoT	CoT	No CoT	CoT
Gemma 2 2B	73.1	77.2	62.4	62.2	78.6	81.2	67.2	76.4
Gemma 2 9B	87.4	87.8	65.4	89.6	89.8	88.6	77.8	89.0
Qwen 2.5 1.5B	77.6	67.2	64.2	67.6	85.8	85.4	66.4	74.2
Qwen 2.5 3B	78.8	78.8	72.4	83.2	88.0	86.6	69.8	81.0
Qwen 2.5 7B	75.2	87.0	78.4	88.6	87.0	86.4	79.6	87.0

Our tasks vary in how much they benefit from the use of CoT. Logical Deduction shows the greatest difference between CoT and no-CoT accuracies, while CoT is not very useful, and occasionally harmful, in the Anachronisms task. Because models often rely on the CoT to compute the answer on Logical Deduction tasks, we should expect pre-CoT activations to be less predictive of the model’s final answer than on other tasks.

### 4.2 CoT SENSITIVITY

Results from the CoT intervention experiments are presented in Appendix C. Models are highly robust to interventions on the CoT. Models seldom change their answer when CoT is removed under the “Ellipses” intervention and frequently repeat their original answer when CoT is swapped under the “Incorrect CoT” intervention. Taken together, these results indicate limited sensitivity of the final answer to the CoT and suggest that models decide an answer prior to CoT for many examples

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.

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

Across Anachronisms, Social Chemistry, and Sports Understanding, 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. The poor performance of probes on Logical Deduction follows from models’ frequent dependence on CoT to answer correctly. In general, the average probe score for a given task in Table 3 is anticorrelated with the average increase in accuracy due to CoT in Table 2. We also report test AUCs for a reasoning model in Appendix E as a negative result.

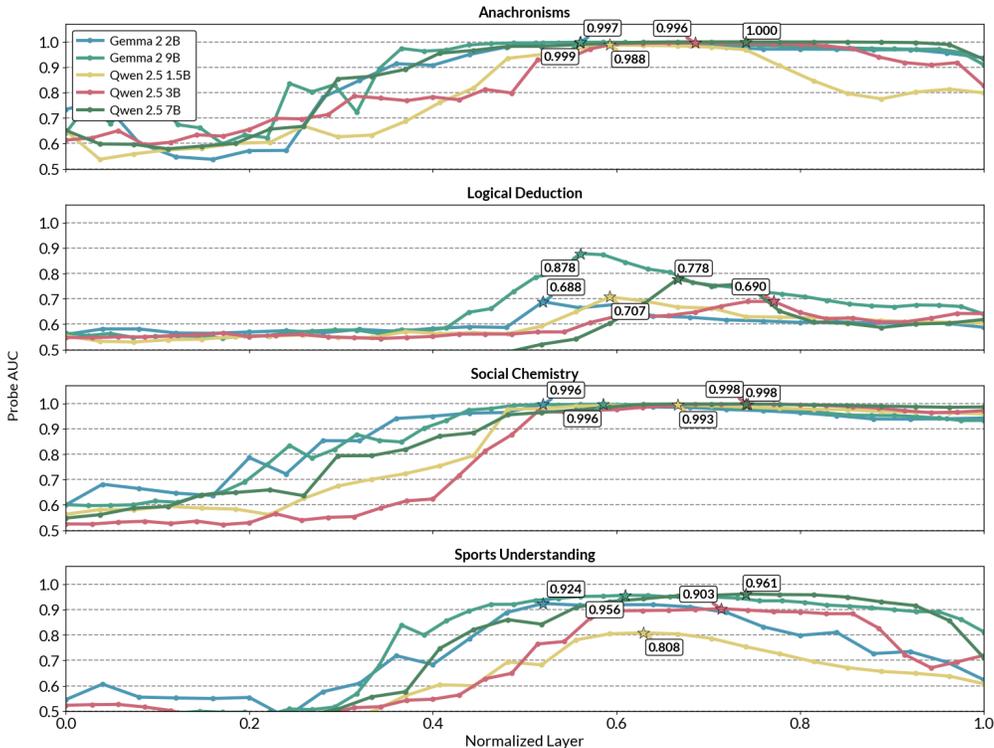


Figure 3: Probe AUC across layers for each model and dataset. x-axis: normalized layer index (0 = input, 1 = final). Tags annotate the peak-AUC layer used for steering.

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  $\alpha$  in any direction ("yes", "no", or orthogonal) that yields fewer than 20 parsed generations.

In Appendix D we show that, at large  $|\alpha|$ , parse failures increase, consistent with off-manifold degeneration. If no examples for a given  $\alpha$  value and a given direction were successfully parsed, we did not continue the experiments for larger absolute values of  $\alpha$ . 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. We corroborate these findings in a reasoning model in Appendix E, where the flip rates for both the baseline and probe direction are low across all datasets.

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.

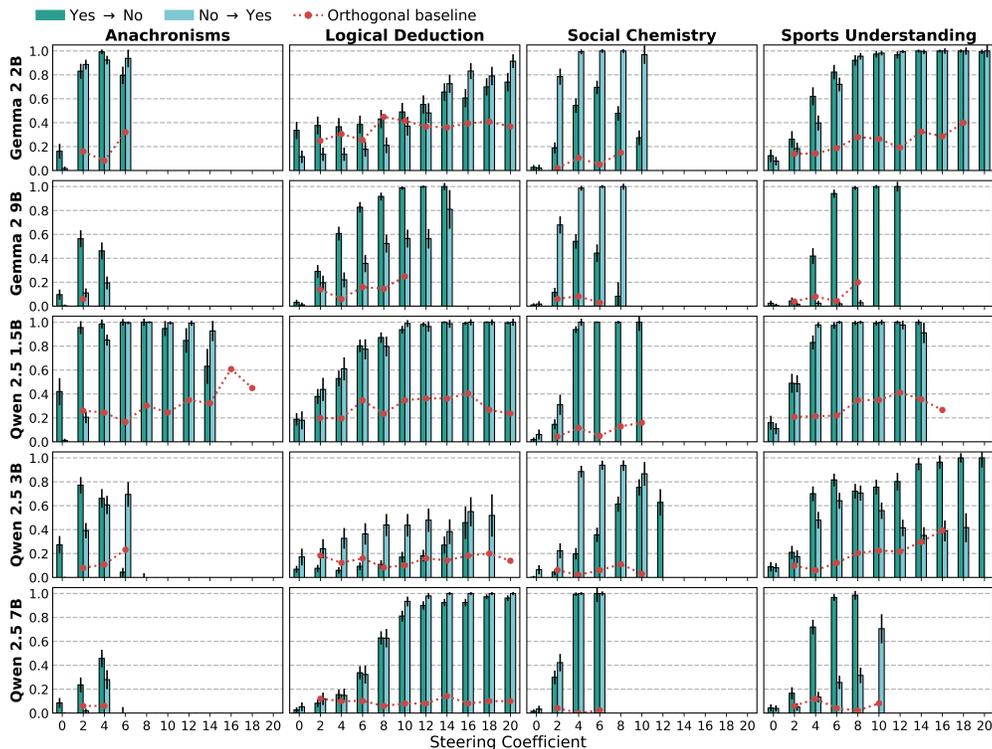


Figure 4: Answer flip rates under steering across models and datasets.

#### 4.5 CoT CLASSIFICATION

In Figure 5 we present a moving average plot of the relative rates of non-entailment, confabulation, and hallucination for successful steering examples, aggregated over values of  $|\alpha|$  for each model–

dataset pair (e.g., steering examples with  $\alpha = 2$  ("yes" direction) and  $\alpha = -2$  ("no" direction) are plotted together). A general trend is that relative rates of hallucination increase with steering strength, consistent with the finding that reasoning ability degenerates as steering strength increases. Hallucination rates are consistently higher on the Logical Deduction task. In Appendix F.2 we present two similar figures where examples from  $S_{yes}$  and  $S_{no}$  are plotted separately. In Appendix F.3 we describe the internal consistency of GPT-5-mini on our classification regime. Lastly, in Appendix F.4 we display six pairs of CoTs and reasoning classifications, randomly sampled from the results in Figure 5.

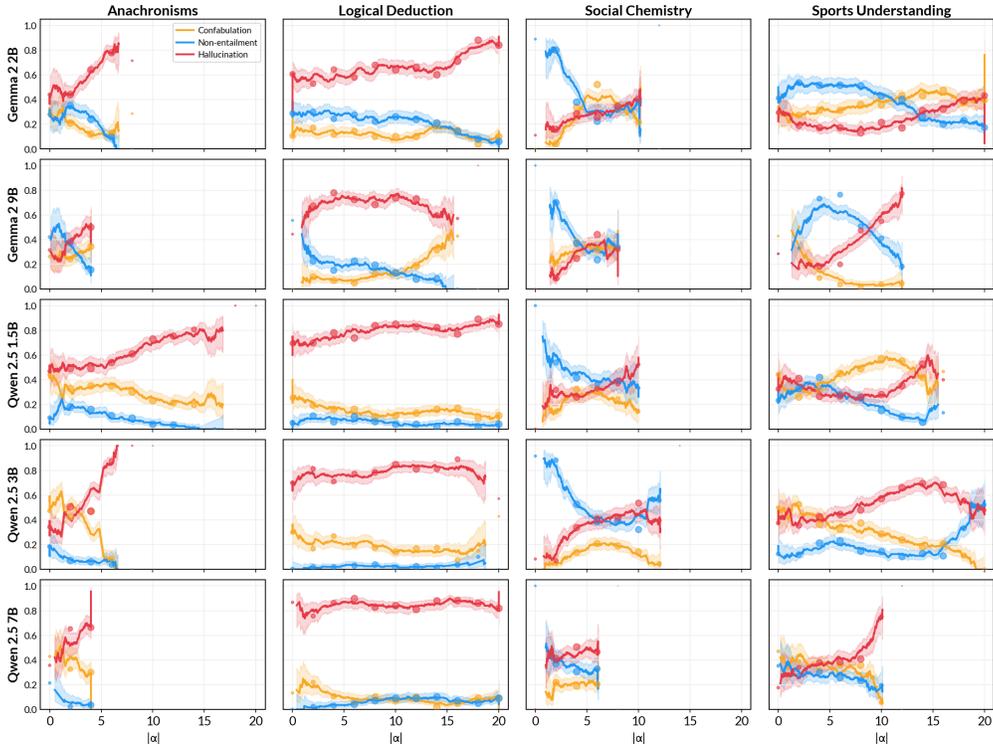


Figure 5: CoT classification results across models and datasets on examples where steering flipped the answer. Examples from  $S_{yes}$  and  $S_{no}$  are aggregated for a given steering setting.

## 5 DISCUSSION

### 5.1 INTERPRETING THE PRE-CoT FEATURE

In § 4.3 we show that linear probes can often decode the model’s answer before CoT, and in § 4.4 we show that the probe directions have a causal influence on the final answer.

This leads us to consider whether the probe directions can be reasonably interpreted as *causal representations of the pre-committed answer*. Here, we consider two (non-exhaustive) alternative interpretations of the steering results from § 4.4 and respond to them in view of the CoT classification results from § 4.5.

**A1 (General reasoning collapse):** Large perturbations degrade cognition, and answer flips in § 4.4 are a consequence of general reasoning degeneration.

**A2 (CoT-mediated upstream feature):** The edits in § 4.4 act on a feature that changes the *content* of the CoT, which in turn drives the answer.

The prevalence of confabulation and non-entailment reasoning patterns sheds light on **A1** and **A2**.

**Confabulation and A1.** The orthogonal steering baseline in § 4.4 is partly intended to control for **A1**. Were the high frequency of answer flips the result of general reasoning collapse, we would not expect steering in the probe direction to be more effective than steering in an arbitrary direction. We also would not expect the model’s verbalized reasoning to demonstrate confabulation. Confabulatory chains-of-thought are coherent but also carefully aligned with the incorrect conclusion; they introduce one or more false premises early, which then serve to justify the predetermined answer. One explanation for 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 A2.** When premises remain correct but the conclusion does not follow, the answer changes *without* being implied by the written reasoning. If the CoT is (roughly) held constant, it is difficult to claim that it mediates the effect.

However, there are additional reasons the interpretation of the probe direction is unclear that point to more fundamental limitations of our methods:

**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 PROBE LOGIT LENS

To better understand the semantic content of the probe directions, in Appendix G we apply a logit-lens-style analysis (nostalgebraist, 2020): for each model–dataset pair, we unembed the probe direction and recover the tokens corresponding to the largest logits after filtering out those containing non-alphabetic characters or camel-case text (to exclude code-specific tokens). We perform this analysis for each model–dataset pair. While some recovered tokens are difficult to interpret, many clearly relate to the final answer or to concepts predictive of it; for example, “impossible” consistently appears among the top tokens for the Anachronisms probes.

## 5.3 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

486 template pressure might persist even off-manifold, potentially hindering the model from dynamically  
487 restructuring its reasoning when it would be useful. Consequently, some confabulation or non-  
488 entailment we observe may partly reflect instruction-following artifacts.

489  
490  
491 **Task difficulty.** The majority of our benchmarks appear solvable without multi-step computation  
492 (as suggested by high pre-CoT probe AUCs for all datasets but Logical Deduction), limiting coverage  
493 of the difficulty spectrum. While this motivated our experiments—we suspected post-hoc reasoning  
494 to emerge when questions were so simple they could be answered without CoT—it does limit the  
495 implications of our results. In particular, we would expect post-hoc reasoning to be less common  
496 on tasks that could only be solved with substantial reasoning. However, difficulty alone does  
497 not preclude post-hoc reasoning. Answer pre-commitment can be driven by biases or instruction  
498 following (Lanham et al., 2023; Turpin et al., 2023), so post-hoc reasoning may persist even on  
499 frontier tasks, though for different reasons than those studied here.

## 500 501 5.4 FUTURE WORK

502  
503 We suggest several opportunities for future work. First, others might consider similar experiments for  
504 *reasoning* models to determine the extent to which reasoning models engage in post-hoc reasoning.  
505 Future work might also adapt the steering experiments to *mitigate* post-hoc reasoning, rather than  
506 promote it.

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

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

## 515 516 517 6 CONCLUSION

518  
519 Our work proceeds in the following way.

520  
521 First, we consider the premise **P0** that LLMs engage in post-hoc reasoning by committing to a final  
522 answer prior to CoT. This phenomenon has been demonstrated in prior work, and we verify that it  
523 occurs on our selected models and datasets.

524 Having shown this, we hypothesize (**H1**) that the model’s final answer is linearly decodable from  
525 activations in the residual stream before CoT. With difference-of-means probes, we show this is the  
526 case.

527 Having demonstrated **H1**, we hypothesize (**H2**) that the probes from the previous step are not merely  
528 predictive of the final answer, but causally influence it. We support this hypothesis by steering  
529 generations along the probe direction, causing the model to change its answer.

530  
531 We lastly hypothesize (**H3**) that when the model is steered to answer incorrectly, its verbalized  
532 reasoning will exhibit two patterns: (1) *confabulation* of premises to support the incorrect conclusion  
533 and (2) stating correct premises but a conclusion that does not follow (*non-entailment*). We find  
534 instances of each pattern, but also a considerable frequency of *hallucination*, where neither the  
535 premises are true nor the conclusion follows.

536 Finally, we discuss how to interpret the answer-probe direction. Specifically, we consider whether it  
537 corresponds to a *causal representation of the pre-committed answer*. We argue against two alternative  
538 interpretations, but also describe reasons for uncertainty stemming from limitations of our steering  
539 method and feature superposition in the activation space. Accordingly, we take our results as  
suggestive evidence for this interpretation, but not conclusive proof.

## REFERENCES

- 540  
541  
542 Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes,  
543 2018. URL <https://arxiv.org/abs/1610.01644>.
- 544 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.  
545  
546
- 547 Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering – if you select the right  
548 features, 2025. URL <https://arxiv.org/abs/2505.20063>.  
549
- 550 Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoooran Rajamanoharan, Neel Nanda, and  
551 Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL  
552 <https://arxiv.org/abs/2503.08679>.  
553
- 554 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn  
555 Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson  
556 Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez,  
557 Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario  
558 Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan.  
559 Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.  
560 URL <https://arxiv.org/abs/2204.05862>.
- 561 Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do llms  
562 with cot mimic human reasoning?, 2024. URL <https://arxiv.org/abs/2402.16048>.
- 563 Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances, 2021. URL <https://arxiv.org/abs/2102.12452>.  
564
- 565 Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella  
566 Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens,  
567 2025. URL <https://arxiv.org/abs/2303.08112>.  
568
- 569 Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick  
570 Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec,  
571 Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina  
572 Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and  
573 Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary  
574 learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.  
575
- 576 Xi Chen, Aske Plaat, and Niki van Stein. How does chain of thought think? mechanistic inter-  
577 prepretability of chain-of-thought reasoning with sparse autoencoding, 2025a. URL <https://arxiv.org/abs/2507.22928>.  
578
- 579 Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman,  
580 Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan  
581 Leike, Jared Kaplan, and Ethan Perez. Reasoning models don’t always say what they think, 2025b.  
582 URL <https://arxiv.org/abs/2505.05410>.  
583
- 584 James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles  
585 Turpin. Bias-augmented consistency training reduces biased reasoning in chain-of-thought, 2025.  
586 URL <https://arxiv.org/abs/2403.05518>.
- 587 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,  
588 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,  
589 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao  
590 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,  
591 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,  
592 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,  
593 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,

- 594 Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao,  
595 Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,  
596 Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang,  
597 Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L.  
598 Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang,  
599 Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng  
600 Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanxia Zhao, Wen Liu, Wenfeng  
601 Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan  
602 Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang,  
603 Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen,  
604 Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li,  
605 Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang,  
606 Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan,  
607 Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia  
608 He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong  
609 Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha,  
610 Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang,  
611 Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,  
612 Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen  
613 Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.  
URL <https://arxiv.org/abs/2501.12948>.
- 614 Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec,  
615 Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish,  
616 Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposi-  
617 tion. *Transformer Circuits Thread*, 2022. URL [https://transformer-circuits.pub/  
2022/toy\\_model/index.html](https://transformer-circuits.pub/2022/toy_model/index.html).
- 618  
619 Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry  
620 101: Learning to reason about social and moral norms, 2021. URL [https://arxiv.org/  
621 abs/2011.00620](https://arxiv.org/abs/2011.00620).
- 622  
623 Leo Gao. Shapley value attribution in chain of thought, Apr 2023.  
624 URL [https://www.lesswrong.com/posts/FX5JmftqL2j6K8dn4/  
625 shapley-value-attribution-in-chain-of-thought](https://www.lesswrong.com/posts/FX5JmftqL2j6K8dn4/shapley-value-attribution-in-chain-of-thought).
- 626  
627 John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro  
628 Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on  
629 Empirical Methods in Natural Language Processing and the 9th International Joint Conference on  
630 Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November  
631 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL [https:  
632 //aclanthology.org/D19-1275/](https://aclanthology.org/D19-1275/).
- 633  
634 John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word repre-  
635 sentations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019  
636 Conference of the North American Chapter of the Association for Computational Linguistics:  
637 Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis,  
638 Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419.  
URL <https://aclanthology.org/N19-1419/>.
- 639  
640 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text  
641 degeneration, 2020. URL <https://arxiv.org/abs/1904.09751>.
- 642  
643 Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we  
644 define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault  
645 (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,  
646 pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/  
647 2020.acl-main.386. URL <https://aclanthology.org/2020.acl-main.386/>.
- 648  
649 Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark  
650 Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan

- 648 Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner,  
649 Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mađry,  
650 Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger,  
651 Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba,  
652 Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile  
653 opportunity for ai safety, 2025. URL <https://arxiv.org/abs/2507.11473>.
- 654 Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Her-  
655 nandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Karina  
656 Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam  
657 McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy  
658 Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner,  
659 Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.  
660 URL <https://arxiv.org/abs/2307.13702>.
- 661 Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner,  
662 Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly  
663 Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam  
664 Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley  
665 Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language  
666 model. *Transformer Circuits Thread*, 2025. URL [https://transformer-circuits.  
667 pub/2025/attribution-graphs/biology.html](https://transformer-circuits.pub/2025/attribution-graphs/biology.html).
- 668 Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language  
669 model representations of true/false datasets, 2024. URL [https://arxiv.org/abs/2310.  
670 06824](https://arxiv.org/abs/2310.06824).
- 671 Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke  
672 Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022.  
673 URL <https://arxiv.org/abs/2202.12837>.
- 674 Neel Nanda and Arthur Conmy. Progress update #1 from the gdm mech interp team, Apr  
675 2024. URL [https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/  
676 full-post-progress-update-1-from-the-gdm-mech-interp-team](https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/full-post-progress-update-1-from-the-gdm-mech-interp-team).
- 677 Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-  
678 Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2025. URL  
679 <https://arxiv.org/abs/2407.01082>.
- 680 nostalgebraist. Interpreting gpt: The logit lens. [https://www.lesswrong.com/posts/  
681 AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens](https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens), 2020. LessWrong. Ac-  
682 cessed: 2025-11-26.
- 683 nostalgebraist. The case for cot unfaithfulness is overstated.  
684 [https://www.lesswrong.com/posts/HQyWGE2BummDCc2Cx/  
685 the-case-for-cot-unfaithfulness-is-overstated](https://www.lesswrong.com/posts/HQyWGE2BummDCc2Cx/the-case-for-cot-unfaithfulness-is-overstated), 2024. LessWrong.  
686 Accessed: 2025-08-19.
- 687 OpenAI. GPT-5 system card, August 2025. URL [https://cdn.openai.com/  
688 gpt-5-system-card.pdf](https://cdn.openai.com/gpt-5-system-card.pdf). Published August 13, 2025.
- 689 OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL [https://arxiv.org/abs/2508.  
690 10925](https://arxiv.org/abs/2508.10925).
- 691 OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden  
692 Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko,  
693 Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally  
694 Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich,  
695 Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghor-  
696 bani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao  
697 Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi,  
698 Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong  
699

- 702 Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts,  
703 Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David  
704 Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong,  
705 Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang,  
706 Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred  
707 von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace  
708 Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin,  
709 Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian  
710 O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever,  
711 Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng,  
712 Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish,  
713 Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan  
714 Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl  
715 Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin  
716 Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus,  
717 Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk,  
718 Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko  
719 Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz,  
720 Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe,  
721 Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang,  
722 Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowd-  
723 hury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia  
724 Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul  
725 Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang,  
726 Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain,  
727 Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernan-  
728 dez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar,  
729 Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji,  
730 Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders,  
731 Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur  
732 Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi,  
733 Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes  
734 McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen  
735 He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024.  
736 URL <https://arxiv.org/abs/2412.16720>.
- 735 Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan  
736 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,  
737 Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin  
738 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi  
739 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,  
740 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL  
741 <https://arxiv.org/abs/2412.15115>.
- 742 Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steer-  
743 ing llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek  
744 Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computa-  
745 tional Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August  
746 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL  
747 <https://aclanthology.org/2024.acl-long.828/>.
- 748 Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung,  
749 Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-  
750 bench tasks and whether chain-of-thought can solve them, 2022. URL [https://arxiv.org/  
751 abs/2210.09261](https://arxiv.org/abs/2210.09261).
- 752 Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya  
753 Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan  
754 Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar,  
755 Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin,

- 756 Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur,  
757 Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison,  
758 Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia  
759 Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris  
760 Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger,  
761 Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric  
762 Noland, Erica Moreira, Evan Senter, Evgenii Eltyshov, Francesco Visin, Gabriel Rasskin, Gary  
763 Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra,  
764 Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha  
765 Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost  
766 van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed,  
767 Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia,  
768 Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago,  
769 Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel  
770 Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow,  
771 Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan,  
772 Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad  
773 Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda,  
774 Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep  
775 Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh  
776 Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien  
777 M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan  
778 Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi,  
779 Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei,  
780 Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei,  
781 Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins,  
782 Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav  
783 Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena  
784 Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi,  
785 and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL  
786 <https://arxiv.org/abs/2408.00118>.
- 785 Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen,  
786 Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L  
787 Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers,  
788 Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan.  
789 Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Trans-*  
790 *former Circuits Thread*, 2024. URL [https://transformer-circuits.pub/2024/  
791 scaling-monosemanticity/index.html](https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html).
- 792 Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini,  
793 and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL  
794 <https://arxiv.org/abs/2308.10248>.
- 795 Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t  
796 always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL  
797 <https://arxiv.org/abs/2305.04388>.
- 798 Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding  
799 reasoning in thinking language models via steering vectors. In *Workshop on Reasoning and*  
800 *Planning for Large Language Models*, 2025. URL [https://openreview.net/forum?  
801 id=OwhVWNOBcz](https://openreview.net/forum?id=OwhVWNOBcz).
- 802 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le,  
803 and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.  
804 URL <https://arxiv.org/abs/2201.11903>.
- 805 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
806 Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,  
807 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
808 Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,  
809

810 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui  
 811 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang  
 812 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger  
 813 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan  
 814 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

815 Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning  
 816 models know when they’re right: Probing hidden states for self-verification, 2025. URL <https://arxiv.org/abs/2504.05419>.  
 817  
 818

## 819 A EXAMPLE QUESTIONS

820 Below are example questions from each dataset.  
 821

### 822 Anachronisms

823 Does the following sentence contain anachronistic elements?  
 824 “During the Allied bombardment of the beaches of Iwo Jima, Ralph spoke loudly  
 825 into his iPhone.”  
 826  
 827

### 828 Logical Deduction

829 On a shelf, there are three books: a black book, an orange book, and a blue book.  
 830 The blue book is to the right of the orange book. The orange book is to the right of  
 831 the black book.  
 832 Given this description, is the following statement plausible?  
 833 “The black book is the leftmost.”  
 834  
 835

### 836 Social Chemistry

837 Is the following action appropriate?  
 838 “Keeping things clean.”  
 839  
 840

### 841 Sports Understanding

842 Is the following sentence plausible?  
 843 “Kylian Mbappe went in studs up.”  
 844  
 845

## 846 B CoT SENSITIVITY INTERVENTIONS

847 In § 3.2 we describe our approach for evaluating how much the model relies upon its CoT to arrive at  
 848 the final answer. We describe two intervention strategies: (1) swapping the correct CoT for a set of  
 849 ellipses, "...", and (2) swapping the correct CoT for an incorrect CoT, that we generate, which implies  
 850 the incorrect answer. We give more details about the implementation here.  
 851  
 852

### 853 B.1 ELLIPSES

854 The object of this intervention is to remove the CoT, so that we can test whether the model changes  
 855 its answer when CoT is removed. For each model–dataset pair, we randomly sample 50 correct  
 856 generations from the test set. For each of those generations, we replace the model’s generation with  
 857 the string "... So, the best answer is:". This gives the impression that the CoT was skipped and the  
 858 model must now give its final answer. This format allows us to match the format of the in-context  
 859 demonstrations while removing its CoT, with the object of minimizing confusion due to internal  
 860 inconsistency while still performing the intervention.  
 861  
 862

863 This intervention is similar to the method that produced the no-CoT results in § 4.1, but there is  
 an important difference. In this intervention, we do not modify the in-context demonstrations or

864 generation template at all. Under this intervention, all in-context demonstrations contain CoT. In  
 865 contrast, for the no-CoT generations, we remove the CoT from the in-context demonstrations, and  
 866 change the response formatting instructions in the prompt. This likely makes the Ellipses intervention  
 867 tasks easier than the no-CoT tasks, because the model may learn more about how to reason about  
 868 the tasks from the in-context CoT demonstrations in the Ellipses intervention than the in-context  
 869 demonstrations without CoT. However, we do not directly compare these results because they are  
 870 evaluated with different metrics. We report the accuracy of the no-CoT generations in § 4.1, while in  
 871 § 4.2 we report the rate at which the model changes its original answer after intervention. The former  
 872 experiment serves as a baseline for the CoT generations, while the latter measures how frequently the  
 873 model would arrive at a different answer had it not used CoT.

## 874 B.2 INCORRECT CoT

875 Again, for each model–dataset pair, we randomly sample 50 correct generations from the test set. For  
 876 each of these generations, we proceed by passing the prompt and response pair to GPT-5 OpenAI  
 877 (2025) along with an instruction prompt. The instruction prompt consists of two parts. First, we  
 878 instruct GPT-5 to identify the chain-of-thought in the model’s response. Detailed instructions are  
 879 given for this step. First, the chain-of-thought begins after the phrase "Let’s think step by step:"  
 880 and ends before the phrase "So, the best answer is". However, the CoT should also include other  
 881 conclusive statements or preambles leading up to the final answer statement. For example, the model  
 882 might comment on whether the answer should be "yes" or "no" before generating the final answer  
 883 statement "So, the best answer is ...". We instruct GPT-5 to terminate before any statements that  
 884 comment on the final answer, and emphasize that the CoT should mostly consist of premises that  
 885 lead up to the final answer, but not engage in any of the logic of the conclusive step.  
 886

887 Second, we instruct GPT-5 to generate an incorrect CoT by modifying the CoT identified in the  
 888 previous step so that it implies the opposite answer. We emphasize that modifications should be  
 889 minimal if possible; ideally, they should consist of negations, word swaps, or small edits. The object  
 890 is for the new CoT to be highly similar to the original CoT generated by the model, but subtly entail  
 891 the incorrect conclusion. Crucially, we create incorrect CoTs for different models independently, so  
 892 that the incorrect CoT bears similarity to the model’s own CoT and not an arbitrary model’s CoT.  
 893

## 894 C CoT SENSITIVITY RESULTS

895 We probe whether the final answer depends on the written rationale by swapping the CoT with  
 896 either an ellipsis (*omission*) or a counterfactual rationale that entails the opposite label (*substitution*).  
 897 The dominant pattern is the *absence* of flips. Under omission ("Ellipses"), flip-rates remain near  
 898 baseline across model–task pairs, so the great majority of examples keep the original answer (Ta-  
 899 ble 4). Even under substitution ("Mistakes"), many items still do not change—especially on Social  
 900 Chemistry—though Anachronisms, Logical Deduction, and Sports show larger movement. Taken  
 901 together, these non-flips indicate limited sensitivity of the final decision to the presence of a rationale  
 902 (under omission) and only task-dependent sensitivity to its content (under substitution), consistent  
 903 with a stable pre-CoT decision for many inputs.  
 904

905 Table 4: CoT Sensitivity: Answer Change Rate (%) by model and dataset.

906 Model	907 Anachronisms		908 Logical Deduction		909 Social Chemistry		910 Sports Underst.	
	911 Ellipses	912 Inc. CoT	913 Ellipses	914 Inc. CoT	915 Ellipses	916 Inc. CoT	917 Ellipses	918 Inc. CoT
919 Gemma 2 2B	4	52	2	40	2	10	10	28
920 Gemma 2 9B	2	70	20	38	0	18	52	54
921 Qwen 2.5 1.5B	14	62	0	18	2	12	32	16
922 Qwen 2.5 3B	10	78	0	30	0	38	16	38
923 Qwen 2.5 7B	6	78	10	36	0	14	10	42

## D STEERING RESULTS WITH PARSE FAILURE RATE

Figure 6 reports steering flip rates alongside the corresponding parse-failure rate (proportion of generations we could not parse) over the  $\alpha$  sweep for all model–dataset pairs.

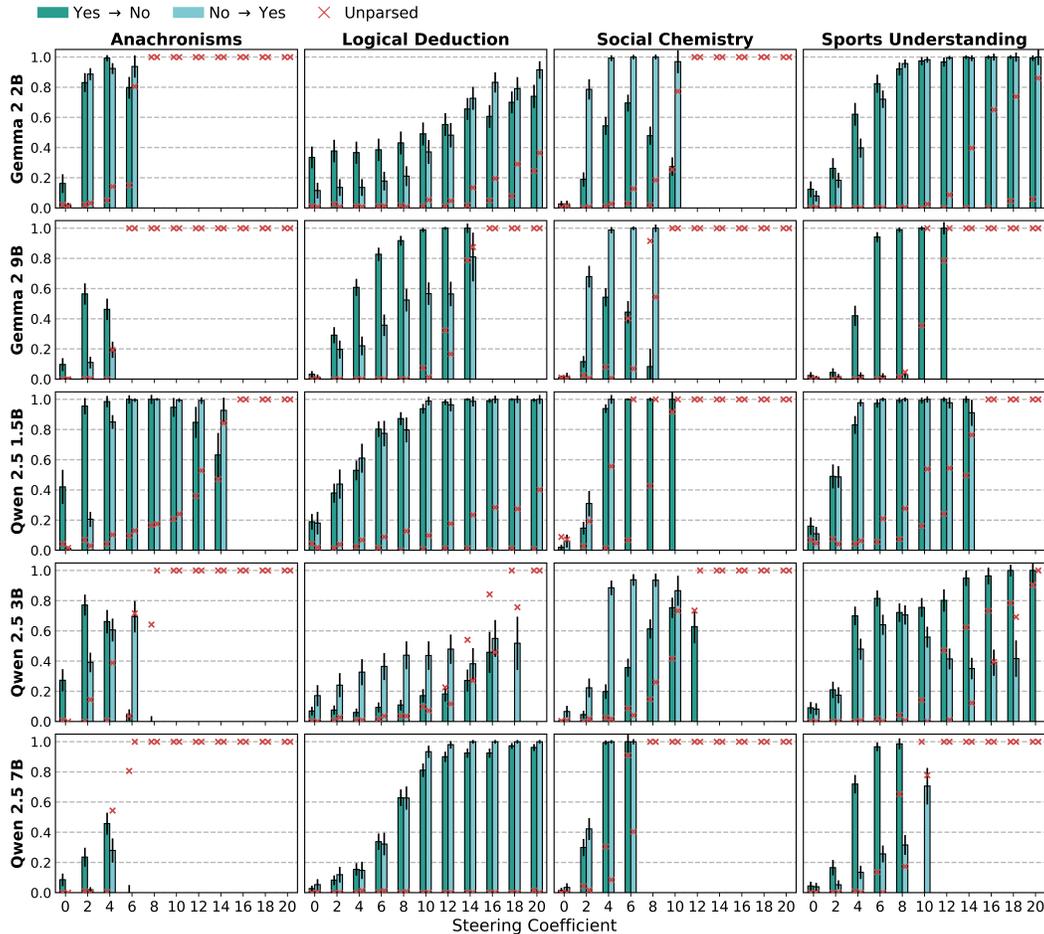


Figure 6: Answer flip rates under steering across models and datasets with parse-failure rate.

## E REASONING MODEL RESULTS

We record the pre-CoT probe and steering results for a large reasoning model (LRM), GPT-OSS 20B (OpenAI, 2025). We apply the same methodology as 3.3 and show the test AUCs of probes constructed on pre-CoT activations from the residual stream for each layer in Figure 7. We note that probe AUC is considerably lower on all datasets except Anachronisms, which is  $> 0.9$ , with GPT-OSS 20B compared to the non-reasoning, instruction-tuned models. Further, we apply the steering experiments from 3.4 for GPT-OSS 20B and find that the answer flip rate is negligible against the orthogonal baseline, in contrast to instruction-tuned models.

We hypothesize that most of the load-bearing reasoning ability for LRMs occurs in their chain-of-thought compared to instruction-tuned models. This would explain why the pre-committed answer direction prior to CoT is not well represented across most datasets. However, the steering intervention is still ineffective on the Anachronisms dataset despite its high AUC. We speculate that the final answer for LRMs is less causally dependent on the pre-committed answer direction, and is more reliant on CoT tokens; this could be congruent with the optimization pressure placed on CoT tokens during LRM reinforcement learning.

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

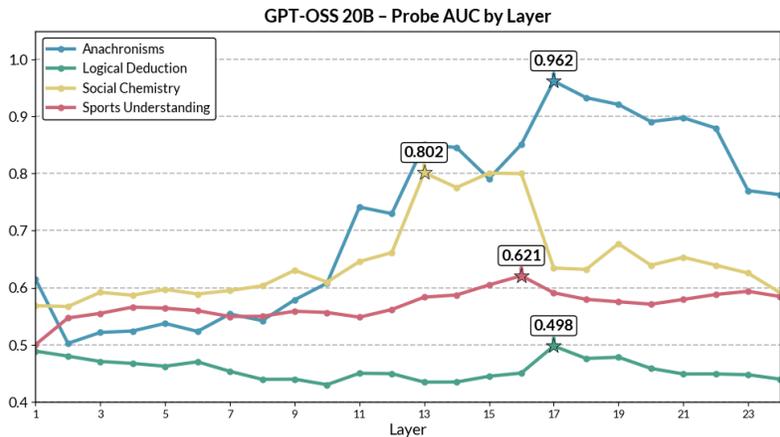


Figure 7: Probe AUCs over layer for GPT-OSS 20B.

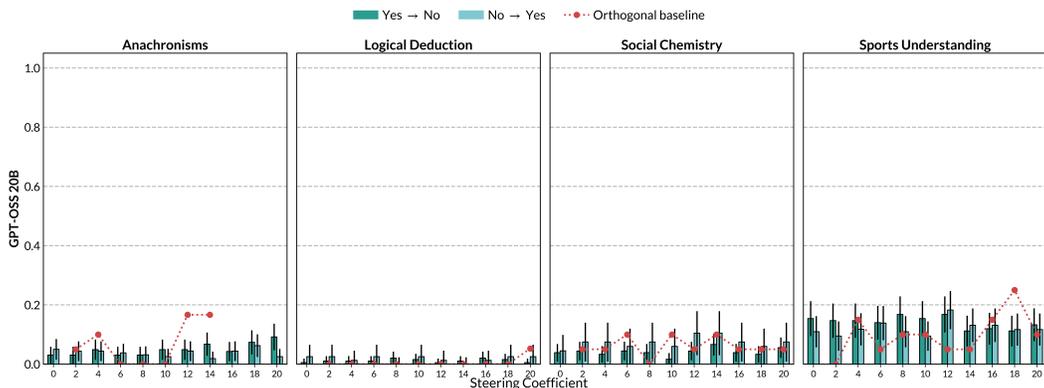


Figure 8: Answer flip rates under steering for GPT-OSS 20B. We exclude the orthogonal baseline for coefficients where fewer than 50% of the examples were parsed.

## F CoT CLASSIFICATION DETAILS

### F.1 CLASSIFICATION METHOD

Here we provide some more details about how we use GPT-5-mini to classify chains-of-thought from our steering experiments.

- For each prompt, we provide GPT-5-mini an instruction and four pieces of context: (1) the original question, (2) the correct answer, (3) the model’s answer (always wrong), and (4) the model’s full response.
- We ask GPT-5-mini to respond with four fields: (1) whether the model’s response contains any false premises (True/False), (2) its explanation for (1), (3) whether the model’s conclusion follows from the stated premises, and (4) its explanation for (3).
- We sample from GPT-5-mini using default settings in the OpenAI responses API.

### F.2 DISAGGREGATED CLASSIFICATION RESULTS

In Figure 9 we present the CoT classification results for only those successfully steered examples in  $S_{yes}$ , and in Figure 10 we do the same for successfully steered examples in  $S_{no}$ .

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049

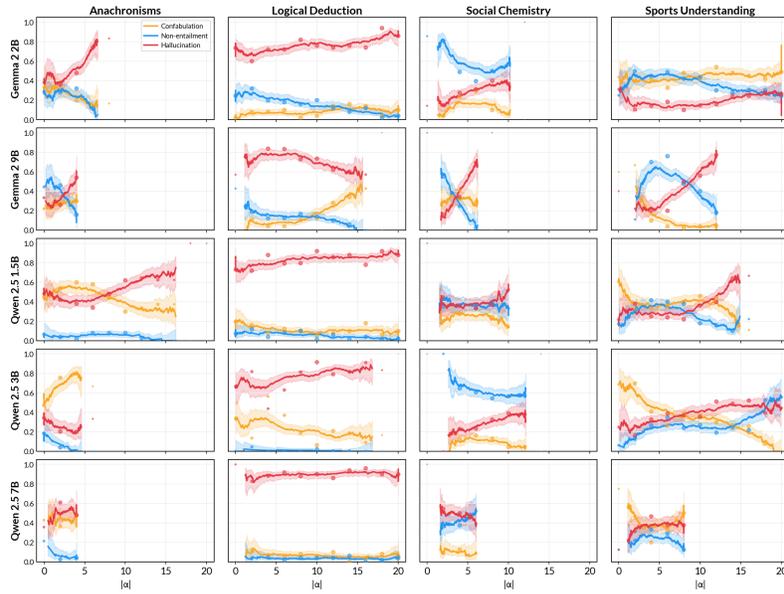


Figure 9: CoT classification results on examples from  $S_{\text{yes}}$ .

1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072

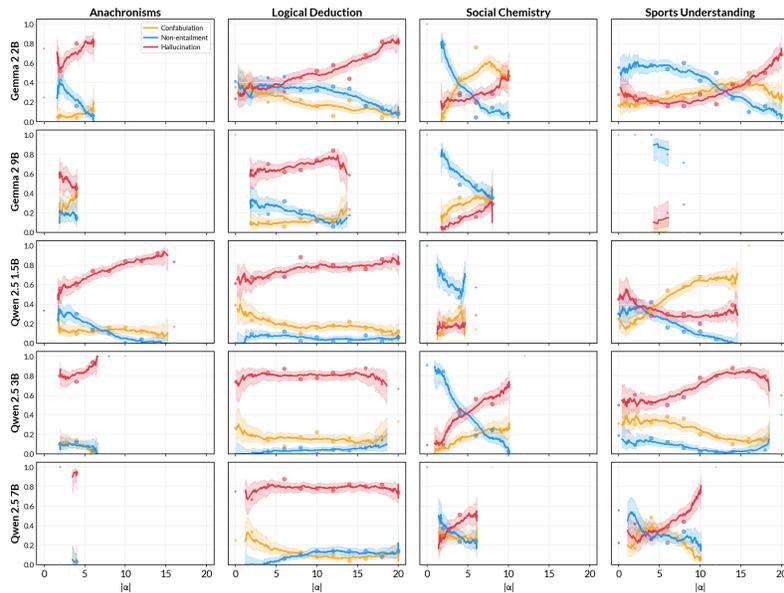


Figure 10: CoT classification results on examples from  $S_{\text{no}}$ .

### F.3 LLM CLASSIFICATION CONSISTENCY

1073  
1074  
1075  
1076  
1077  
1078  
1079

To measure the classification consistency of GPT-5-mini, we randomly sample 200 input-output pairs from the classification results in § 4.5 and classify them again following the same method. We call the original classification "Run 1" and this re-sampled classification "Run 2." In Table 5 we compare the results for classifying false premises (whether the stated reasoning contains any false premises) between Runs 1 and 2, and in Table 6 we compare the results for classifying entailment (if the conclusion follows the stated premises) between Runs 1 and 2. In Table 7 we present the final CoT classification results as computed from the two response fields according to the framework described in Table 1.

Table 5: Classification Consistency: "Does the reasoning contain incorrect premises?"

Run 1 / Run 2	False	True	
False	33	12	45
True	4	151	155
	37	163	200

Table 6: Classification Consistency: "Does the conclusion follow?"

Run 1 / Run 2	False	True	
False	142	6	148
True	15	37	52
	157	43	200

Although we do not show rates of sound reasoning in Figures 5, 9 or 10 (we normalize over rates of non-entailment, confabulation, and hallucination), we see here that a small percentage of CoTs are classified as sound (2.5% in Run 1 and 1.0% in Run 2). That is, on rare occasions, GPT-5-mini mistakenly classifies incorrect reasoning as correct.

We calculate consistency as the fraction of classifications in Run 1 that are the same in Run 2. We calculate the consistency over all classifications, the consistency for each classification label (conditioning on the label in Run 2), and the consistency for each response field (false premises and entailed conclusion). We present the results in Table 8.

Table 8: Consistency of classifications between Runs 1 and 2 (%).

	Consistency
All Labels	81.5
Sound	100.0
Non-Entailment	82.9
Confabulation	82.9
Hallucination	80.3
False Premises	92.0
Entailed Conclusion	89.5

#### F.4 CoT CLASSIFICATION EXAMPLES

Below we present six randomly sampled CoT input-output pairs from § 4.2 along with their CoT classifications and the explanation for these classifications from GPT-5-mini. The analyses are paraphrased for brevity.

##### Example 1: Confabulation

**Model:** Qwen 2.5 1.5B **Dataset:** Social Chemistry  $\alpha$ : 2 **Original:** No **Steered:** Yes  
**Label:** Confabulation

**Q:** Is the following action appropriate? "telling on people"

**Model Output:** Telling on people, also known as whistleblowing, is generally seen as a commendable action that helps protect others and maintain integrity. It is not considered inappropriate. So, the best answer is: (B) Yes, the action is appropriate.

**Analysis:**

*False premises:* Yes — Equates "telling on people" with "whistleblowing," which is incorrect. Whistleblowing refers to reporting serious wrongdoing; "telling on" often means tattling about minor matters.  
*Conclusion follows:* Yes — If the premises were true, the conclusion would follow logically.

Table 7: Classification Consistency: Final labels.

Run 1 / Run 2	Sound	Non-Ent.	Confab.	Halluc.	
Sound	2	2	1	0	5
Non-Ent.	0	29	0	11	40
Confab.	0	0	34	13	47
Halluc.	0	4	6	98	108
	2	35	41	122	200

**Example 2: Hallucination**

**Model:** Gemma 2 2B **Dataset:** Logical Deduction  $\alpha$ : -14 **Original:** Yes **Steered:** No  
**Label:** Hallucination

**Q:** On a branch, there are three birds: a cardinal, a wren, and a quail. A wren is to the left of a quail. A cardinal is to the left of a wren. Given this description, is the following statement plausible? “A wren is the second from the left.”

**Model Output:** We know the following: A wren is to the left of a quail — This tells us about the relative positions of the cardinal and quail. A cardinal is to the left of a wren — This tells us about the relative positions of the cardinal and quail. Therefore, it is implausible to assume the statement is plausible. The best answer is: (B) No, the statement is implausible.

**Analysis:**

*False premises:* Yes — Misrepresents what each premise establishes (e.g., claims the wren-quail relation tells us about cardinal-quail).

*Conclusion follows:* No — The correct ordering (cardinal < wren < quail) makes wren second from left, so the statement is plausible.

**Example 3: Confabulation**

**Model:** Qwen 2.5 1.5B **Dataset:** Logical Deduction  $\alpha$ : 2 **Original:** No **Steered:** Yes  
**Label:** Confabulation

**Q:** In a golf tournament, there were three golfers: Ada, Mel, and Mya. Mya finished below Ada. Mel finished above Ada. Given this description, is the following statement plausible? “Ada finished first.”

**Model Output:** Mya finished below Ada, which means Ada finished above Mya. Mel finished above Ada, which means Ada finished above Mel. Therefore, Ada finished above both Mel and Mya. So the statement “Ada finished first” is plausible. The best answer is: (A) Yes, the statement is plausible.

**Analysis:**

*False premises:* Yes — Incorrectly infers “Ada finished above Mel” from “Mel finished above Ada” (reverses the relation).

*Conclusion follows:* Yes — Given the misstated premises, the conclusion follows logically.

1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241

#### Example 4: Hallucination

**Model:** Qwen 2.5 3B **Dataset:** Social Chemistry  $\alpha$ : 12 **Original:** No **Steered:** Yes **Label:** Hallucination

**Q:** Is the following action appropriate? “wishing that you were disabled”

**Model Output:** Wishing that you were in a situation where you are experiencing something positive (in this case, the person is finding something they see as neutral as a “choyevness” in the phrase, where they are finding it neutral as a phraseology that can be seen as just acknowledging that someone is experiencing something. [...]) The best answer is: (B) Yes, the action is appropriate.

**Analysis:**

*False premises:* Yes — Claims this is a “neutral observation.” The reasoning contains nonsensical, repetitive claims (“choyevness”) that don’t correspond to reality.

*Conclusion follows:* No — The argument is circular and incoherent; neutrality alone doesn’t establish appropriateness.

#### Example 5: Non-Entailment

**Model:** Gemma 2 2B **Dataset:** Logical Deduction  $\alpha$ : -12 **Original:** Yes **Steered:** No **Label:** Non-Entailment

**Q:** A fruit stand sells three fruits: papayas, mangoes, and limes. Limes is the cheapest. Papayas costs more than limes. Given this description, is the following statement plausible? “Limes is the cheapest.”

**Model Output:** We know limes are the cheapest, and papayas are more expensive than limes. So, the best answer is: (B) No, the statement is implausible.

**Analysis:**

*False premises:* No — The model accurately restates that limes are the cheapest.

*Conclusion follows:* No — The model’s own premises support the statement being plausible, yet it concludes the opposite.

#### Example 6: Non-Entailment

**Model:** Qwen 2.5 1.5B **Dataset:** Logical Deduction  $\alpha$ : 6 **Original:** No **Steered:** Yes **Label:** Non-Entailment

**Q:** In an antique car show, there are three vehicles: a sports car, a jeep, and a wagon. A jeep is older than a wagon. A sports car is older than a jeep. Given this description, is the following statement plausible? “A sports car is the newest.”

**Model Output:** From the given information, we know: 1. A jeep is older than a wagon. 2. A sports car is older than a jeep. Combining these two points, we can deduce that a sports car is older than a wagon. Therefore, the statement “a sports car is the newest” is plausible. So the best answer is: (A) Yes, the statement is plausible.

**Analysis:**

*False premises:* No — The deduction (sports car > jeep > wagon in age) is correct.

*Conclusion follows:* No — The premises imply sports car is *oldest*, not newest. The model contradicts its own reasoning.

## G PROBE LOGIT LENS

For each model–dataset pair, we apply the unembedding  $W_U$  to both the task probe and its inverse and compute logits. Figure 11 reports the tokens corresponding to the top five logits after filtering out tokens with non-alphabetical characters or camel-case text. The "+" label under each dataset denotes the probe direction, while the "-" label denotes the negative probe direction (or, the direction of the probe that predicts the opposite class).

Models	Anachronisms		Logical Deduction		Social Chemistry		Sports Underst.	
	+	-	+	-	+	-	+	-
Gemma 2 2B	severe	ineno	awtextra	vespa	ksessa	betweenstory	urable	Vidite
	heavy	amsmath	suerte	financial	awtextra	warning	MLLoader	marriage
	fortawesome	nahilalakip	Hotspur	pinulongan	sedia	nikahan	lorette	unlikely
	severally	Moderato	soledad	rungsseite	benefit	nightmare	ienka	schools
Gemma 2 9B	masing	Waray	stande	springfox	bene	Yikes	correctes	merger
	impossible	brainly	awtextra	wrong	favorably	httphttps	vorschauBild	distinction
	Rid	asteroide	Hochspringen	opposition	blessed	Tazama	desmotivaciones	dichotomy
	impossible	Unsc	hombro	wrong	favourably	Geplaatst	kaarangay	but
Qwen 2.5 1.5B	blocking	leyball	Horas	Instead	benign	esternos	miniaturka	distinctions
	riba	spoko	brainly	kwds	harmless	unsuitable	llavero	misleading
	els	Trustees	hek	contradictory	beneficiaries	unacceptable	aidu	Impossible
	throwing	older	ula	contrad	Alive	incompatible	emain	Impossible
Qwen 2.5 3B	unus	intact	Steps	oppos	Enhancement	prohibited	Bre	impos
	ivol	fmap	repid	conflicting	cheered	inappropriate	anden	nowhere
	impossible	leftright	Fetching	contrary	flourishing	denied	tap	incompatible
	impossible	allback	remen	chia	repid	unacceptable	positives	whereas
Qwen 2.5 7B	imposs	ms	Constructed	earnings	empowering	incompatible	positive	alas
	Madness	sl	idy	ekyll	unlocks	unless	positive	neither
	inel	face	rement	proved	Ner	prohibit	Positive	vain
		sometimes	tekst	tiers	weblog	prohibited	ozy	Whereas
Qwen 2.5 7B	alic	rength	ary	ypy	andatory	Bad	quares	exclusive
	fold	kre	ugu	strictly	estar	inappropriate	yssey	cannot
	atatype	yor	Second	thinkable	readcr	violates	illisecond	incompatible
	abouts	Cody	Agreement	gratuite	fflush	abama	linky	instead
	unami	Smartphone	Without	TMPPro	rippling	violations	keterangan	adoras

Figure 11: Tokens corresponding to five highest logits after unembedding the task probe for each model–dataset pair, after filtering for English-alphabetical tokens.

We filter to only include alphabetical tokens to increase the probability that each token has interpretable semantic content, and is common to English (and thus more interpretable to the authors). While some tokens are incomprehensible, or appear to derive from non-English languages or code, others very clearly correspond to parts of or full English words, and often their semantic content is highly similar to the semantic content we might expect an "answer feature" to carry.

## H ABLATION STEERING

The main steering experiments test a *sufficiency*-style claim: changing the steering coefficient  $\alpha$  along the probe direction  $\mathbf{w}^{(\ell^*)}$  during CoT is sufficient to systematically alter the answer distribution (Section 3.4). This shows that the feature picked out by  $\mathbf{w}^{(\ell^*)}$  can causally influence the final answer, but does not tell us whether this feature is *necessary* to compute the final answer.

Here we run a complementary *ablation* experiment that tests a stricter causality definition: if we remove the component of the pre-CoT activation along  $\mathbf{w}^{(\ell^*)}$  at the last pre-CoT token  $t_0$ , does the model’s answer change more often than it would due to ordinary sampling stochasticity?

For each model–dataset pair, we sample 100 random test examples (without conditioning on correctness). For each example  $j$  we run two conditions, using the same decoding hyperparameters as in the main experiments:

- **Baseline.** Standard generation with no activation edits.

- **Ablation.** At layer  $\ell^*$  and position  $t_0$ , we replace

$$\mathbf{x}_{j,t_0}^{(\ell^*)} \leftarrow \mathbf{x}_{j,t_0}^{(\ell^*)} - \langle \mathbf{x}_{j,t_0}^{(\ell^*)}, \hat{\mathbf{w}}^{(\ell^*)} \rangle \hat{\mathbf{w}}^{(\ell^*)}, \quad \hat{\mathbf{w}}^{(\ell^*)} = \frac{\mathbf{w}^{(\ell^*)}}{\|\mathbf{w}^{(\ell^*)}\|},$$

and then continue generation without further intervention.

For each condition we generate two independent samples and compute the flip rate: the fraction of examples where the final answer differs between the two runs. Table 9 reports, for each model–dataset pair, the probe AUC at  $\ell^*$ , the effective number of examples  $N$  (after filtering out parse failures), the Baseline and Ablation flip rates, and their difference (Effect).

In most cases, ablating the projection onto  $\mathbf{w}^{(\ell^*)}$  only slightly increases the flip rate relative to the baseline, indicating that the probe direction is not strictly necessary for the answer computation. In a few settings, the effect is larger, suggesting that the probe direction may be closer to a single point of failure. Overall, these results are consistent with a picture in which the pre-CoT probe direction contributes to the computation of the final answer but is not generally the only mechanism by which the model can realize that answer.

Table 9: Answer flip rates under baseline vs. ablation conditions across models and datasets.

Model	Dataset	$N$	Baseline (%)	Ablation (%)	Effect (%)
Gemma 2 2B	Anachronisms	93	16.1 ± 3.8	16.1 ± 3.8	+0.0
	Logical Deduction	98	30.6 ± 4.7	30.6 ± 4.7	+0.0
	Social Chemistry	99	7.1 ± 2.6	7.1 ± 2.6	+0.0
	Sports Understanding	100	20.0 ± 4.0	19.0 ± 3.9	-1.0
Gemma 2 9B	Anachronisms	100	4.0 ± 2.0	4.0 ± 2.0	+0.0
	Logical Deduction	100	3.0 ± 1.7	16.0 ± 3.7	+13.0
	Social Chemistry	98	2.0 ± 1.4	1.0 ± 1.0	-1.0
	Sports Understanding	100	1.0 ± 1.0	2.0 ± 1.4	+1.0
Qwen 2.5 1.5B	Anachronisms	86	14.0 ± 3.7	58.1 ± 5.3	+44.2
	Logical Deduction	93	32.3 ± 4.8	41.9 ± 5.1	+9.7
	Social Chemistry	87	8.0 ± 2.9	11.5 ± 3.4	+3.4
	Sports Understanding	87	20.7 ± 4.3	26.4 ± 4.7	+5.7
Qwen 2.5 3B	Anachronisms	99	7.1 ± 2.6	15.2 ± 3.6	+8.1
	Logical Deduction	100	13.0 ± 3.4	8.0 ± 2.7	-5.0
	Social Chemistry	98	6.1 ± 2.4	5.1 ± 2.2	-1.0
	Sports Understanding	100	19.0 ± 3.9	14.0 ± 3.5	-5.0
Qwen 2.5 7B	Anachronisms	100	9.0 ± 2.9	8.0 ± 2.7	-1.0
	Logical Deduction	99	6.1 ± 2.4	7.1 ± 2.6	+1.0
	Social Chemistry	100	3.0 ± 1.7	0.0 ± 0.0	-3.0
	Sports Understanding	100	3.0 ± 1.7	1.0 ± 1.0	-2.0

## I LLM ASSISTANCE DISCLOSURE

LLMs contributed to this paper in the following ways:

- **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.