

Compliance vs. Sensibility: On the Reasoning Controllability in Large Language Models

Xingwei Tan^α Marco Valentino^α Mahmud Elahi Akhter^β

Yuxiang Zhou^{β ω} Maria Liakata^{β γ} Nikolaos Aletras^α

^αUniversity of Sheffield ^βQueen Mary University of London

^γThe Alan Turing Institute ^ωUfonia

{xingwei.tan,m.valentino,n.aletras}@sheffield.ac.uk

{m.akhter,yuxiang.zhou,m.liakata}@qmul.ac.uk

Abstract

Large Language Models (LLMs) acquire reasoning capabilities through shared inference patterns in pre-training data, which are further elicited via Chain-of-Thought (CoT). However, whether fundamental reasoning patterns, such as induction, deduction, and abduction, can be decoupled from specific problem instances remains a critical challenge for model controllability. In this paper, we present the first systematic investigation of this problem through the lens of *reasoning conflicts*, an explicit tension between parametric and contextual information induced by mandating logical schemata that deviate from those expected for a target task. Our evaluation reveals that LLMs consistently prioritize *sensibility* over *compliance*, favoring task-appropriate reasoning patterns despite conflicting instructions. We further demonstrate that reasoning conflicts are internally detectable, as confidence scores drop during conflicting episodes. Probing experiments confirm that reasoning types are linearly encoded from middle-to-late layers, indicating the potential for activation-level controllability. Leveraging these insights, we steer models towards compliance, increasing instruction following by up to 29%. Overall, our findings establish that while LLM reasoning is anchored to concrete instances, active mechanistic interventions can effectively decouple logical schemata from data, offering a path toward improved controllability, faithfulness, and generalizability.¹

1 Introduction

Recent work has shown that Large Language Models (LLMs) acquire fundamental reasoning capabilities by leveraging shared inference patterns and procedural knowledge within their pre-training data (Ruis et al., 2025). These capabilities, further elicited via Chain-of-Thought (CoT) prompting, offer remarkable performance on logical reasoning

benchmarks (Wei et al., 2022; Kojima et al., 2022; Sprague et al., 2025). By prompting the model to generate a step-by-step sequence of intermediate thoughts, CoT enables LLMs to decompose complex problems and justify their final answer.

However, it is unclear whether the inference patterns employed during CoT derive from generalizable, disentangled reasoning capabilities or remain intrinsically tied to the specific tasks and instances observed during pre-training. Standard CoT typically leaves the choice of the actual reasoning mechanism entirely up to the model. The LLM acts as a black box in deciding whether to deploy deductive, inductive, or abductive reasoning (Walton, 2001). If it naturally selects the correct reasoning strategy for the task, this autonomy is highly effective. But the lack of explicit structural controllability raises robustness and faithfulness issues. If a user accidentally applies the wrong prompt template, or malevolently injects an adversarial instruction forcing a suboptimal logical path, it is not clear how models react. When a human is handed a problem and told to use an inappropriate reasoning schema such as deduction for an ambiguous abductive mystery, they rely on meta-reasoning and cognitive flexibility to handle the conflict (Ackerman and Thompson, 2017; Diamond, 2013). Humans continuously monitor their own reasoning processes (Nelson, 1990; Lemaire and Siegler, 1995), hence they might recognize the mismatch and switch to the task-appropriate reasoning, or attempt to follow the request, realize it is failing, then revise it mid-thought. It is currently unknown how LLMs navigate this exact tension, leading to a fundamental, yet underexplored, research question: *How robust and controllable is LLM reasoning when forced into a conflict between instruction compliance and logical sensibility?*

To answer this question, we first prepare a testbed consisting of logical reasoning questions that only require one of the fundamental reasoning

¹The experimental code can be found in: https://github.com/Xingwei-Tan/compliance_sensibility

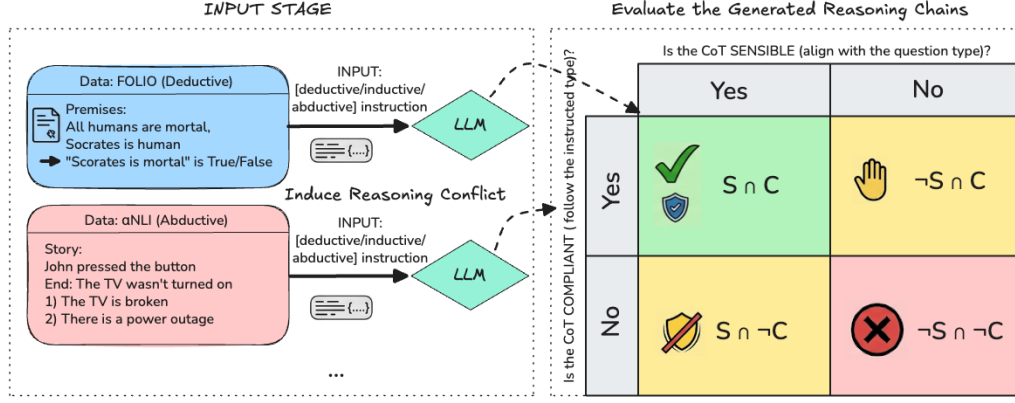


Figure 1: Reasoning instructions are used to induce reasoning conflicts. Then, we evaluate the sensibility and compliance of the responses.

types: deduction, induction, and abduction (Table 1). Then, we evaluate how LLMs react to instructions that align or conflict with the task-appropriate reasoning. Our experiments reveal that a reasoning conflict between compliance (i.e., following the instruction) and sensibility (i.e., task-appropriate reasoning) indeed exists: models often face a tension between a user’s reasoning instructions and their own internal priors. To understand this phenomenon, we provide a mechanistic analysis via probing to demonstrate that LLMs successfully encode and understand reasoning instructions in their middle-to-late layers. We show that the decision to comply or diverge is an active process that begins mid-computation. Finally, we use Contrastive Activation Addition (CAA) (Rimsky et al., 2024) to showcase that reasoning compliance can be augmented during inference. Our contributions are as follows:

- **Comprehensive Systematic Evaluation:** We conduct the first rigorous assessment of reasoning conflicts across three open-source model families and two frontier LLMs on four logical reasoning datasets, presenting a comprehensive evaluation of how LLMs direct their reasoning on diverse logical questions.
- **Mechanistic Analysis of Reasoning Prompting:** We provide the first in-depth investigation into the interplay between a user’s reasoning instructions and an LLMs’ own internal priors for reasoning. Furthermore, we quantify instruction compliance under varying reasoning types and instruction settings.
- **Enhancing Reasoning Compliance via Steering:** We explore the efficacy of steer-

ing mechanisms in augmenting model compliance which shows steering can enhance the reasoning type compliance by up to 29%.

2 Related Work

2.1 Fundamental Logical Reasoning in LLMs

Although LLMs have demonstrated remarkable reasoning capabilities on math, such as GSM8K (Cobbe et al., 2021), they still face various issues in logical reasoning or soft reasoning (Sprague et al., 2025; Liu et al., 2025b; Sun and Saparov, 2025). Recent work has shown that there is no one-size-fits-all reasoning strategy for LLMs, and they need to reason adaptively based on the context (Zheng et al., 2025). Xu et al. (2025) evaluate the fundamental reasoning paradigms and reveal a substantial performance gap: models frequently fail to identify the correct reasoning direction, particularly in abductive and mixed-form tasks. Liu et al. (2025a) find similar gaps persist in inherently reasoning models such as DeepSeek-R1 (Guo et al., 2025) and Qwen3 (Yang et al., 2025). This struggle is further evidenced in real-world claim verification, where even frontier LLMs fail on abductive reasoning questions (Dougrez-Lewis et al., 2025). Recent efforts examine the structural interplay between the reasoning types, and investigate how different fundamental reasoning paradigms induce generalization or enhance LLMs’ reasoning (Hu et al., 2025; Cao et al., 2026).

2.2 Chain-of-Thought Faithfulness

A growing body of evidence suggests that CoT frequently fails to faithfully describe the model’s internal reasoning process (Turpin et al., 2023; Chen et al., 2025). This discrepancy is often character-

ized as a lack of causal dependence, where the generated rationale serves as a post-hoc justification rather than a functional roadmap to the final answer (Lanham et al., 2023; Siegel et al., 2024). Recent studies on soft-reasoning tasks indicate that while CoT may be unfaithful, it can still impose an active guidance effect on LLMs’ output (Lewis-Lim et al., 2025), even if that guidance does not follow a direct causal link (Tutek et al., 2025).

Our work is related to injecting misleading cues to detect unfaithfulness (Turpin et al., 2023; Chua and Evans, 2025), but we pivot the focus toward *reasoning compliance*. While prior research has investigated whether models verbalize known causal features (Bao et al., 2025), it remains unclear how the verbalization relates to a model’s ability to adhere to a mandated logical schema. By exploring how CoT influences predictions when a model either acknowledges or unfaithfully omits an explicit structural cue, we move beyond simple error-testing (Lanham et al., 2023) to investigate the fundamental reasoning conflicts between user instructions and internal model priors. This is particularly important for high-stake domains, such as the legal domain, where users would like full transparency on the reasoning process and full control on the reasoning schema (Kelsall et al., 2026).

2.3 Mechanistic Interpretation and Model Steering for Logical Reasoning

Recent mechanistic interpretation studies suggest that instruction-following and task-specific logic are often localized within residual stream activations, typically maturing in the middle-to-late layers of the transformer architecture (Belrose et al., 2025; Nanda et al., 2023). To bridge the gap between understanding these representations and controlling them, model editing is a lightweight and flexible alternative to fine-tuning (Hughes et al., 2026). While early applications focused on mitigating toxicity or enhancing safety, recent work applies these methods to reasoning. For instance, Valentino et al. (2026) adopt Conditional Activation Steering (Lee et al., 2025) to mitigate content effects. While prior steering research often targets binary attributes, we test CAA (Rimsky et al., 2024) to the more abstract domain of reasoning compliance by sampling responses on our reasoning conflict testbed.

3 Methodology

3.1 Problem Formulation

We aim to measure how *controllable* is LLM reasoning when forced into a conflict between *instruction compliance* and *logical sensibility* in logical reasoning problems. We denote q as a question that requires logical reasoning and a as its correct answer. We assume q is associated with an a priori known suitable reasoning type t , which is predetermined by the dataset’s creators. t can be deductive, inductive, or abductive reasoning. An instruction g_t (defined in §3.2) is formulated that includes a mandated fundamental reasoning type t' for guiding the LLMs to reason and answer the question. These instructions are presented as part of the in-context prompt that will be fed into the LLMs. The answer generation process is formulated as:

$$P(\hat{r}, \hat{a}) = LLM(q, g_{t'}), t' \in \{De., In., Ab.\} \quad (1)$$

where $\hat{a}$ is the generated final answer, and $\hat{r}$ is the reasoning chain generated by CoT. Finally, we automatically infer the reasoning type $\hat{t}$ used in $\hat{r}$, and compare it with the mandated and suitable reasoning types, t' and t respectively. When $\hat{t} = t'$, the model is **compliant** (C). When $\hat{t} = t$, the model is **sensible** (S). In some of the experiment settings, we intentionally prompt with $g_{t'}, t' \neq t$ to induce a reasoning conflict. Although part of a long-form generation may contain a secondary reasoning type, our formulation operates under a *Dominant Reasoning Mode* assumption to build a controlled testbed for analyzing our research question. We define $\hat{t}$ as the most prominent reasoning type that contributes directly to the final answer. All questions that we include in the experiments are specifically designed to be answered with only one reasoning type.

3.2 Prompting for Reasoning Controllability

As prompting is the most accessible way a user would interact with an LLM, we first test reasoning controllability in-context via prompting. We want to quantify how well LLMs understand fundamental reasoning types, and how well they can perform them based on basic instructions. For that purpose, we construct zero-shot prompts g_t to control reasoning by showing LLMs the high level definition of the available reasoning types, and the general steps to perform them. Table 1 shows the prompts for deductive, inductive and abductive reasoning.

Type	Prompt
Deductive	You are a logician tasked with performing deductive reasoning. You are given a general rule and specific observations. Your task is to apply the general rule to the observations to derive a logically certain conclusion. Provide a detailed reasoning process leading to your conclusion.
Inductive	You are a logician tasked with performing inductive reasoning. You are given a set of observations. Your task is to infer the most probable general rule that explains all observations. Then, use the inferred rule to make predictions about new observations. Provide a detailed reasoning process leading to your conclusion.
Abductive	You are a logician tasked with performing abductive reasoning. You are given a general rule and an observation. Your task is to generate the most probable and simplest hypothesis that, if true, would logically explain all the observations provided. Provide a detailed reasoning process leading to your conclusion.

Table 1: Prompts for controlling LLMs’ reasoning.

3.3 Automated Reasoning Type Classification

Given a generated CoT $\hat{r}$, we classify the reasoning type t' employed by the model as deductive, inductive or abductive. Since this is challenging and expensive for manual annotation, as models often produce verbose and structurally ambiguous outputs, we use the LLM-as-a-Judge framework (Zheng et al., 2023).

We present the generated CoT to a frontier model using a zero-shot instruction: *What type of reasoning is the following process?* To formalize the judge’s output, we employ a rule-based parser that maps responses into four categories: Deductive, Inductive, Abductive, or Direct (no intermediate reasoning). The Cohen’s κ agreement between humans and GPT-5.1 is 0.84. Notably, we find that frontier models are highly sensitive to lexical mimicking, where an LLM superficially echoes terms from the reasoning instructions (e.g., “logically certain” for deduction or “most probable” for abduction) without executing the corresponding logical operations (see §5.4). The details of judge verification can be found in Appendix D, where we also show the confusion matrices of the labels.

Finally, based on the comparison between the mandated and generated CoT reasoning type, each response is further categorized into the following four categories:

- **Sensible and Compliant ($S \cap C$):** The model adopts the instructed reasoning type, which happens to be the task-appropriate one.
- **Sensible but Non-Compliant ($S \cap \neg C$):** The model ignores the reasoning instruction and defaults to the task-appropriate type.

- **Compliant but Non-Sensible ($\neg S \cap C$):** The model follows the instruction even when it is logically ill-suited for the task (e.g., using induction for a deductive logic puzzle).
- **Neither Sensible nor Compliant ($\neg S \cap \neg C$):** The model uses an unrelated reasoning type.

3.4 Steering for Reasoning Controllability

We adopt model steering to enhance the controllability of LLMs by guiding them towards compliance with the reasoning instructions t' . We use CAA (Rimsky et al., 2024) to identify the parameters of the LLM that are responsible for reasoning. CAA modifies internal activations during inference without updating model parameters. It effectively enables fine-grained control of sycophancy, refusal, and instruction-following behaviors (Stolfo et al., 2025). Given multiple pairs of contrastive responses (r_p, r_n) of the question q (target behavior and the opposite), a steering vector $\mathbf{v}$ is generated by computing the Mean Difference of the residual stream activations at a specific layer.

$$\mathbf{v}_{MD} = \frac{1}{|D|} \sum_{(q, r_p, r_n) \in D} (h_L(q, r_p) - h_L(q, r_n)), \quad (2)$$

where $h_L(\cdot)$ is the activation at layer L and D the training dataset. During inference, the steering vector is added to the corresponding layer at every token position. A multiplier (coefficient) μ controls the intensity and direction of the effect.

To obtain a steering vector of instruction compliance, we collect compliant and non-compliant responses as contrastive pairs by sampling reasoning responses with the prompts in Table 1 on the training datasets with 3 random seeds. Then, the LLM judge introduced in §3.3 determines whether the responses are compliant. To make sure that only the reasoning compliance is contrasted, we marginalize the effects of final answer correctness and instructed type by only pairing those responses with the same correctness and instructed type.

4 Experimental Setup

4.1 Models

We evaluate two frontier proprietary models: GPT-5.1 (OpenAI, 2025) and GEMINI3-FLASH (Gemini, 2025). Three famils of state-of-the-art open-weight LLMs are also included in our experiments: OLMO (Olmo, 2025), QWEN, and LLAMA. We compare smaller versions (7B/8B), larger version

(32B/70B), instruct version, and thinking version of these open-weight LLMs. For QWEN, we include the results with thinking mode on and off.

4.2 Data

We evaluate the LLMs across four datasets specifically selected to represent the fundamental reasoning types: deduction, induction, and abduction.

FOLIO (Deduction): A first-order logic dataset containing textual premises and claims (Han et al., 2024). Models must determine the truth value of a claim based strictly on the provided logical constraints. We utilize FOLIO as a benchmark for deductive reasoning, where conclusions must follow necessarily from established premises.

SPR (Induction): The *Sequence Pattern Recognition* (SPR) task requires models to predict the subsequent element in a numerical sequence governed by recurring arithmetic operations (Hu et al., 2025). This task serves as a proxy for inductive reasoning, requiring the model to generalize a latent rule from observed patterns.

α NLI (Abduction): The *Abductive Natural Language Inference* (α NLI) dataset presents a story’s beginning and ending, requiring the model to select the most plausible intermediate hypothesis from two options (Zhao et al., 2023). This represents abductive reasoning, which is about inference to the best explanation.

RECV (Deduction+Abduction): *Reasoning in Evidence-based Claim Verification* (RECV) dataset consists of real-world claim verification tasks sourced from Wikipedia and Twitter (Douguez-Lewis et al., 2025). Unlike the other benchmarks, every RECV question is associated with a human-annotated reasoning type, providing a diverse testbed for reasoning compliance.

4.3 Evaluation

Task Accuracy. We extract the final answer from the generated output via a rule-based parser that matches the special tags we put in the prompt template (Appendix C, Table 3). Then, the answer is compared against the gold answer to compute the final answer accuracy.

Sensibility and Compliance. We classify CoT $\hat{r}$ with the LLM judge introduced in §3.3, and compare against t and t' to acquire the sensibility and compliance rates respectively (see §3.1). We use GPT-5.1 API with 0.1 temperature as a judge for models other than GPT-5.1, then use GEMINI3-FLASH to judge GPT-5.1.

Confidence Estimation. As textual CoT may be unfaithful, it is necessary to also analyze signals beyond textual generation. We start with measuring the confidence for generating answer a' to quantify how the reasoning conflict impact the LLM’s internal states. $P(\text{True})$ (Kadavath et al., 2022) let an LLM generate a response with the original prompt (the reasoning instructions in our case), then ask a follow-up question in the same conversation: Is the answer (A) *True* or (B) *False*. The confidence score can thus be extracted as the probability of generating the token A. The full prompt details are shown in Appendix C.

Probing for Reasoning. To examine whether reasoning instructions are encoded in the model’s internal states, we apply linear probing to the residual stream activations of the open LLMs. For each response generated under the reasoning instructions, we collect hidden states from every transformer layer after the feed-forward block, restrict the representation to the question span, and obtain one vector per layer by mean-pooling over the span tokens. On top of these frozen representations, we train independent multi-class linear probes to predict: 1) the instructed reasoning type, 2) the judge-inferred reasoning type, and 3) the reasoning is compliant or not. Probe performance is measured layer-wise using accuracy and macro-F1, and we summarize each model family by the strongest probe layer and by the full layer-wise curves.

5 Results

5.1 Revealing Difficulty in Controllability

LLMs prioritize logical sensibility over instruction compliance. As illustrated in Fig. 2, our aggregated results across all datasets, all instructed reasoning types, and random seeds reveal a clear hierarchy: LLMs consistently prioritize logical sensibility over instruction compliance. On average, models opt for the reasoning type that matches the task requirements in 43.5% of cases ($S \cap \neg C$). In contrast, models exhibit total compliance by following the instructed type at the expense of task appropriateness in only 18.6% of instances ($\neg S \cap C$). We find that stray reasoning, where the model utilizes a framework that is neither sensible nor compliant, is notably rare, occurring in only 5.7% of the samples. This “Neither” category consistently accounts for less than 10% of total outputs across all tested models. Furthermore, we observe a correlation with model scale: larger models demonstrate a

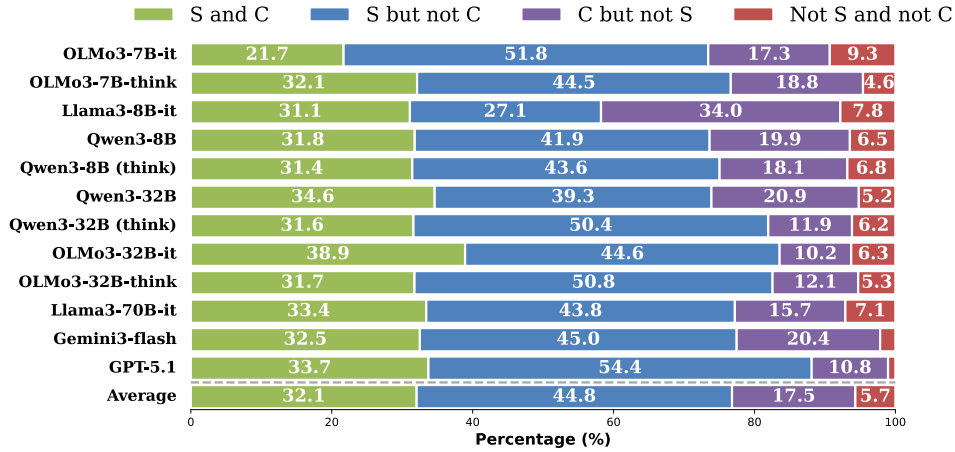


Figure 2: Whether the reasoning is sensible (S) or compliant (C) based on the LLM judge.

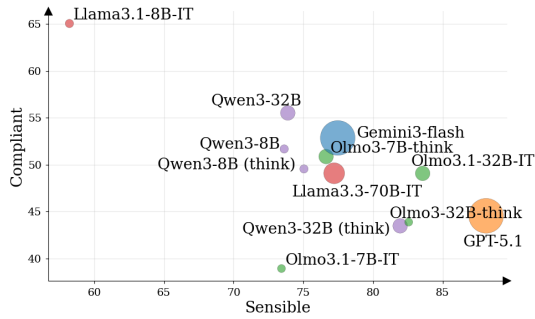


Figure 3: Proportion of sensible vs. compliant CoT across models. Larger circles represent larger LLMs.

higher degree of “logical grounding”, reducing the frequency of outputs that fall into the non-sensible, non-compliant quadrant.

Reasoning sensibility scales with model sizes, with notable exceptions. As shown in Fig. 3, reasoning sensibility generally improves with model scale. GPT-5.1 exhibits the highest sensibility score (88.1%), whereas LLAMA3.1-8B-IT is the least sensible (58.2%). However, this scaling trend is not universal. Most notably, the QWEN3 with thinking mode off demonstrates nearly identical sensibility rates (73.5% vs. 73.9%), despite QWEN3-32B possessing a four-fold increase in parameters over the 8B variant. Moreover, GEMINI3-FLASH underperforms relative to its parameter class, trailing behind LLAMA3.3-70B-IT and OLMO3.1-32B-IT. These results suggest that while increased sizes typically fosters more sensible reasoning, architectural choices and specific instruction-tuning recipes play an essential role in determining a model’s logical grounding.

Instruction compliance is family-dependent and parameter-agnostic. Our results (Fig. 3) indicate that compliance rates diverge greatly

across model families. Model scale serving as an unreliable predictor of instruction compliance. LLAMA3.1-8B-IT achieves a substantially higher compliance rate (65.1%) than all other evaluated models, while OLMO3-7B-IT exhibits the lowest (39%). Within the QWEN3 and OLMO3 lineages, we observe standard scaling, where the 32B variants outperform their 8B/7B counterparts in compliance. Conversely, the LLAMA family demonstrates an inverse trend: LLAMA3.1-8B-IT is notably more compliant than the much larger LLAMA3.3-70B-IT. This high variance across families and the non-monotonic relationship with model size suggest that compliance is primarily driven by specific post-training methodologies rather than their scales.

Models with build-in thinking are generally more sensible and less compliant. Based on Fig. 2 and 3, QWEN models have higher sensible rate when turning the thinking mode on. On the other hand, switching thinking mode on makes the compliant rate lower. The larger thinking OLMO also has a lower compliant rate. This indicate a substantial behavior difference between the build-in thinking and the CoT outside of the thinking tags.

Logical sensibility is the primary determinant of task accuracy. As illustrated in Fig. 4, LLMs consistently achieve higher accuracy when their reasoning is sensible, regardless of instruction compliance (i.e., the sensible/compliant and sensible/non-compliant cohorts). Due to its limited frequency, we exclude the fourth category (not sensible/not compliant) from this analysis as it is not sufficient for statistical significance testing (Fig. 2). This disparity is most pronounced on the FO-LIO dataset, where sensible CoT maintains a 20%–

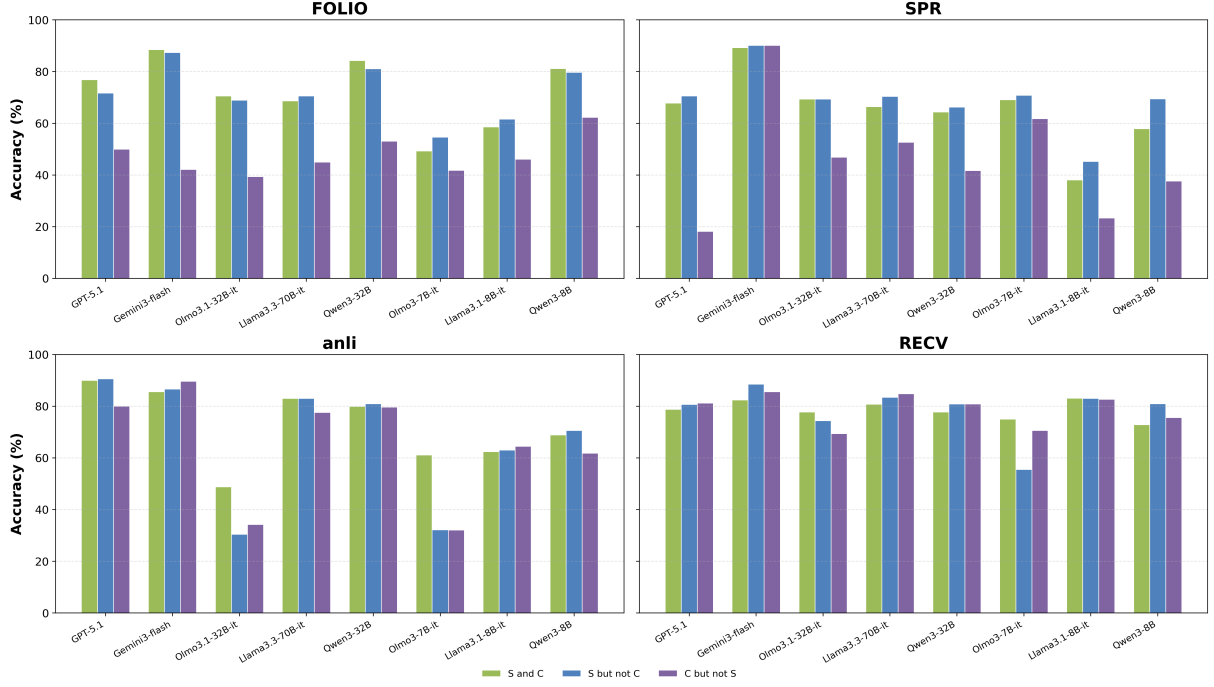


Figure 4: The average accuracies of the final answers with respect to the categories.

40% accuracy lead over the compliant/not sensible category. Conversely, the RECV dataset exhibits marginal gaps, with differences across all categories falling below 20%. At the model level, GEMINI3-FLASH displays relatively uniform accuracy across the three cohorts, with the notable exception of its compliant/not sensible performance on FOLIO. This suggests that GEMINI3-FLASH may rely more heavily on internal heuristics, with its generated CoT serving primarily as post-hoc justification rather than functional reasoning steps. In Appendix G, we show the numeric accuracies in Table 6, and the accuracies grouped by the instructed reasoning types in Table 7.

5.2 Confidence Reflects Internal Reasoning Conflicts

Instruction compliance is positively correlated with model confidence. To look beyond the textual reasoning chains induced by our instructions, we investigate the relationship between reasoning compliance and the internal confidence of the LLMs. We quantify confidence using the $p(\text{True})$ probability of the final answer, aggregated across the six open-weight instruction LLMs over three instructions and three random seeds per dataset. To isolate the impact of compliance from the correlation between accuracy and confidence, we marginalize the effect of correctness by grouping results into correct and incorrect cohorts (Table 2).

Correctness	Compliant	Not Compliant	Δ	p value
Incorrect	0.5664	0.4831	0.0833	3e-14
Correct	0.7116	0.6937	0.0179	5e-13

Table 2: Average $p(\text{True})$ confidence scores of compliant reasoning and not compliant reasoning grouped by incorrect or correct final answers.

Our analysis reveals that $p(\text{True})$ confidence of the compliant responses is significantly higher than the non-compliant ones ($p < 0.05$), regardless of the final answer’s correctness. However, the effect size (Cohen’s d) is 0.09 on the correct responses and 0.12 on the incorrect responses. These results indicate that reasoning conflicts do not cause catastrophic drop in model confidence, but they introduce a subtle frictional drag that is mathematically verifiable during the response generation. The significance in the confidence drop during non-compliant episodes still implies that models maintain an internal awareness of the provided instructions, which imposes an implicit influence on the generation process even when the model ultimately defaults to a different reasoning paradigm. This motivates us to further conduct mechanistic analysis of the internal activation.

5.3 Mechanistic Analysis

Fig. 5 shows reasoning compliance is moderately decodable from the model’s hidden states, with decoding accuracy typically peaking in the middle

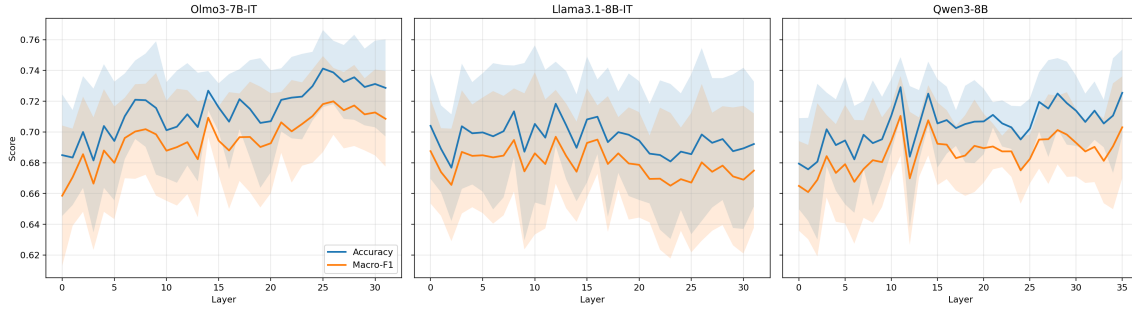


Figure 5: The probing scores across layers for complicant vs. non-compliant binary classification.

and late layers for OLMO and QWEN. This suggests that the signal for compliance strengthens as the forward pass progresses. In our preliminary experiment, we also probe the instructed reasoning type and judge-inferred reasoning type (Appendix E). Compared to them, compliance is a weaker and less linearly accessible property of the internal state. These results show that although the LLMs robustly encode the instructed reasoning types, whether they eventually comply with it is a weaker and more entangled property of the internal state.

Based on the probing results (Fig. 5 and Appendix E), we can identify the middle layers as the primary site for encoding reasoning paradigms, we applied CAA to layers 14–17 of OLMO3-7B-IT. This model was specifically chosen for intervention due to its baseline status as the least compliant model in our study (Fig. 3). As shown in Fig. 6a, applying steering vectors with a positive multiplier (μ) greatly increases the compliance rate, with the effect scaling alongside μ . This is particularly evident when forcing *inductive* and *deductive* instructions on αNLI questions, where steering successfully overrides the model’s natural abductive inclination. However, the impact on task accuracy (Fig. 6b) reveals a complex trade-off. For abductive instructions, accuracy drops as μ increases; qualitative analysis suggests this is due to a “task-neglect” side effect, where the model becomes so focused on the steered reasoning type that it fails to follow formatting constraints (e.g., enclosing the final answer in the required tags). Interestingly, for *inductive* and *deductive* instructions, accuracy follows a non-monotonic trend: initially declining before reaching an upward peak as compliance is enforced. Finally, we observe that steering with a negative μ consistently improves accuracy across all instruction types. Such “anti-steering” also makes LLMs resort to direct answer (Fig. 6c), potentially reducing the cognitive friction caused by conflicting

reasoning. The steering results on FOLIO (Fig. 7) also show a generally similar trend, where the compliance rate of deductive instruction is substantially increased. Moreover, the accuracies on all three instructed reasoning type are greatly improved with larger μ . More steering results on other datasets and models are in Appendix F.

Example 1

To determine the validity of the conclusion “Montana is home to the city of Missoula” based on the given premises, let’s analyze each premise step by step and apply abductive reasoning to find the most probable and simplest hypothesis that explains all observations. [...] Given the direct statement in premise 2 that “The state of Montana includes the cities of Butte, Helena, and Missoula”, we have explicit information supporting the conclusion that Montana is home to the city of Missoula. The other premises either provide additional context that does not contradict this conclusion or introduce information about other locations that do not directly impact the conclusion about Missoula. Therefore, based on the premises provided, the conclusion that “Montana is home to the city of Missoula” is supported directly by premise 2 and is not contradicted by any other premise.

5.4 Case Study

Example 1 shows the output of LLAMA3.3-70B-IT when explicitly instructed to use abductive reasoning. Notably, the model exhibits a form of lexical mirroring: it incorporates key abductive terminology from the prompt (e.g., “most probable” and “simplest hypothesis”) to signal compliance. However, the internal logic later pivot to deductive logical entailment to reach the final answer. This suggests that while the model knows the requested reasoning is inappropriate for the task, its true reasoning process is masked behind a compliant textual facade. Our LLM judge correctly identifies the reasoning type as *deductive*, despite the model’s self-proclamation of abduction.

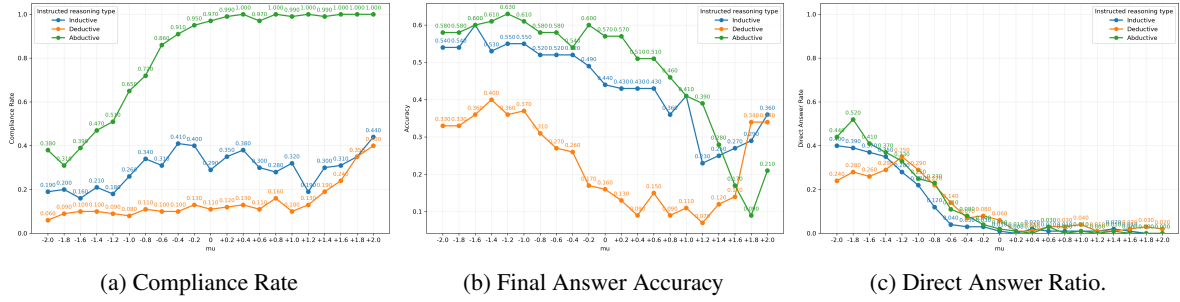


Figure 6: The impact of multiplier μ on α -NLI when steering the layer 14-17 of OLM3-7B-IT.

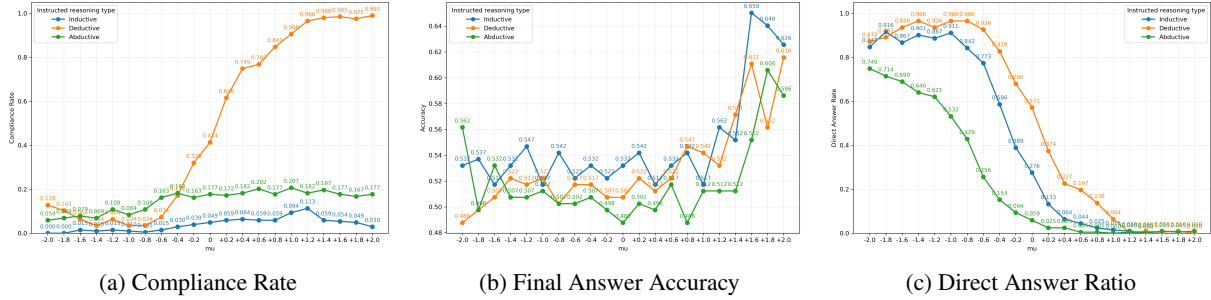


Figure 7: The impact of multiplier μ on FOLIO when steering the layer 14-17 of OLM3-7B-IT.

6 Conclusion

We characterized the reasoning conflict in LLMs, showing that while they successfully encode reasoning instructions in middle-to-late layers, they often struggle to prioritize compliance. Our systematic evaluation reveals that this tension is an active, mid-computation process that varies significantly by model scale and architecture. We also showed that this conflict is not an immutable limitation; rather, it can be mitigated through activation steering, which enhances instruction compliance by up to 29%. This study provides both a mechanistic explanation for reasoning failures and a lightweight solution for enforcing reasoning compliance, establishing a critical foundation for developing more robust and controllable reasoning engines for high-stakes applications. A promising direction for future work is to extend the analysis into questions that involve multiple reasoning types.

Limitations

While CAA proves capable of steering model compliance, it does so at a visible cost to generation quality. Our intervention experiments are primarily validated on localized open-weights architectures where strong activation steering frequently induced formatting degradation or task neglect. Future research is required to evaluate whether scaling laws or dynamic, layer-specific steering vectors can iso-

late reasoning compliance without destabilizing the model’s core textual coherence.

Acknowledgments

XT, MA, YZ, ML and NA are supported by the EPSRC [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) (McIntosh-Smith et al., 2024). Isambard-AI is operated by the University of Bristol and is funded by the UK Government’s Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. The authors also acknowledge IT Services at the University of Sheffield and the University of Oxford Advanced Research Computing for the provision of HPC services.

References

- Rakefet Ackerman and Valerie A Thompson. 2017. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8):607–617.
- Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. 2025. How likely do LLMs with CoT mimic human reasoning? In *Proceedings of the 31st International Conference on*

- Computational Linguistics*, pages 7831–7850, Abu Dhabi, UAE. Association for Computational Linguistics.
- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. [Eliciting latent predictions from transformers with the tuned lens](#). *Preprint*, arXiv:2303.08112.
- Mingzi Cao, Xingwei Tan, Mahmud Elahi Akhter, Marco Valentino, Maria Liakata, Xi Wang, and Nikolaos Aletras. 2026. [Fundamental reasoning paradigms induce out-of-domain generalization in language models](#). *Preprint*, arXiv:2602.08658.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, John Schulman, Arushi Somani, Carson Denison, Peter Hase, Misha Wagner, Fabien Roger, and Vlad Mikuli. 2025. [Reasoning models don’t always say what they think](#).
- James Chua and Owain Evans. 2025. [Are deepseek r1 and other reasoning models more faithful?](#) *Preprint*, arXiv:2501.08156.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Adele Diamond. 2013. Executive functions. *Annual Review of Psychology*, 64:135–168.
- John Dougrez-Lewis, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, and Maria Liakata. 2025. [Assessing the reasoning capabilities of LLMs in the context of evidence-based claim verification](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20604–20628, Vienna, Austria. Association for Computational Linguistics.
- Team Gemini. 2025. [Gemini 3 flash: frontier intelligence built for speed](#).
- Daya Guo, Dejian Yang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenyuan Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. [FOLIO: Natural language reasoning with first-order logic](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.
- Zhiyuan Hu, Yibo Wang, Hanze Dong, Yuhui Xu, Amrita Saha, Caiming Xiong, Bryan Hooi, and Junnan Li. 2025. [Beyond ‘aha!’: Toward systematic meta-abilities alignment in large reasoning models](#). *Preprint*, arXiv:2505.10554.
- Anthony Hughes, Vasisht Duddu, N. Asokan, Nikolaos Aletras, and Ning Ma. 2026. [PATCH: Mitigating PII leakage in language models with privacy-aware targeted circuit Patching](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5139–5153, Rabat, Morocco. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, and 1 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Joshua Kelsall, Xingwei Tan, Aislinn Bergin, Jiahong Chen, Maria Waheed, Tom Sorell, Rob Procter, Maria Liakata, Jenny Chim, and Serene Chi. 2026. A rapid evidence review of evaluation techniques for large language models in legal use cases: trends, gaps, and recommendations for future research. *AI & society*, 41(4):4025–4043.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *Preprint*, arXiv:2307.13702.
- Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehl, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. [Programming refusal with conditional activation steering](#). In *The Thirteenth International Conference on Learning Representations*.
- Patrick Lemaire and Robert S Siegler. 1995. Four aspects of strategic change: Contributions to children’s learning of multiplication. *Journal of Experimental Psychology: General*, 124(1).
- Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025. [Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29826–29841, Suzhou, China. Association for Computational Linguistics.

- Hanmeng Liu, Yiran Ding, Zhizhang Fu, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025a. [Evaluating the logical reasoning abilities of large reasoning models](#). *Preprint*, arXiv:2505.11854.
- Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025b. [Logical reasoning in large language models: A survey](#). *Preprint*, arXiv:2502.09100.
- Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. 2024. [Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence](#). *Preprint*, arXiv:2410.11199.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). In *The Eleventh International Conference on Learning Representations*.
- Thomas O Nelson. 1990. Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26:125–173.
- Team Olmo. 2025. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- OpenAI. 2025. [Gpt-5 system card](#).
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Laura Ruis, Maximilian Mozes, Juhan Bae, Sidhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. 2025. [Procedural knowledge in pretraining drives reasoning in large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. [The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–546, Bangkok, Thailand. Association for Computational Linguistics.
- Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. [To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. [Improving instruction-following in language models through activation steering](#). In *The Thirteenth International Conference on Learning Representations*.
- Yunxin Sun and Abulhair Saparov. 2025. [Language models do not follow occam’s razor: A benchmark for inductive and abductive reasoning](#). *Preprint*, arXiv:2509.03345.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. [Measuring faithfulness of chains of thought by unlearning reasoning steps](#). *Preprint*, arXiv:2502.14829.
- Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2026. [Mitigating content effects on reasoning in language models through fine-grained activation steering](#). *Preprint*, arXiv:2505.12189.
- Douglas Walton. 2001. Abductive, presumptive and plausible arguments. *Informal Logic*, 21(2).
- Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuan-sheng Ni, Guozhou Zheng, and Huajun Chen. 2024. [EasyEdit: An easy-to-use knowledge editing framework for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2025. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*.
- An Yang, Anfeng Li, and 1 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.
- Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. [Abductive commonsense reasoning exploiting mutually exclusive explanations](#). In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

Tianshi Zheng, Cheng Jiayang, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Wong, and Simon See. 2025. [LogiDynamics: Unraveling the dynamics of inductive, abductive and deductive logical inferences in LLM reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20721–20742, Suzhou, China. Association for Computational Linguistics.

A Implementation Details

We use vLLM (Kwon et al., 2023) on NVIDIA GH200 (96G), A100 (80G), and AMD MI300X (192G) for generating the responses with open LLMs. The inference temperature is set to 0.5. We sample responses from LLMs with 3 different random seeds. For QWEN3-8B and QWEN3-32B, the thinking mode is disable. We use *EasyEdit 2.0* for model steering (Wang et al., 2024; Zhang et al., 2024). For probing, we use a default logistic regression classifier with $C = 1.0$ and L2 penalty.

B Dataset Details

For Sequence Pattern Recognition (SPR), we use the code of Hu et al. (2025) to generate a training set of 4,300 samples and a test set of 100 samples. We randomly split RECV (Dougrez-Lewis et al., 2025) by stratifying the labels in to train, dev, and test sets. Each subset (i.e., climate fever, VitC, and PhemePlus) has the exact same number of samples. This results in 99 test samples and 1,302 train samples. We randomly sample 100 questions from the test set of α NLI to have a balanced number of test samples as the other datasets. We use the training set of FOLIO, α NLI, and SPR for generating the steering vectors. For all the evaluation, we use the test split of α NLI, SPR, and RECV, and the validation split of FOLIO as it does not have a test split available.

C Prompt Details

This section presents the complete prompts used in our evaluation pipeline. Table 3 show the full prompts for inducing reasoning conflicts, including the special tags we ask the models to add for the answer matching.

Below is the prompt for $p(\text{True})$ confidence estimation:

$p(\text{True})$ Prompt

User: [Question + Reasoning Instruction]
 LLM: [CoT + Final Answer]
 User:
 Is this answer:
 (A) True
 (B) False
 The answer is: (

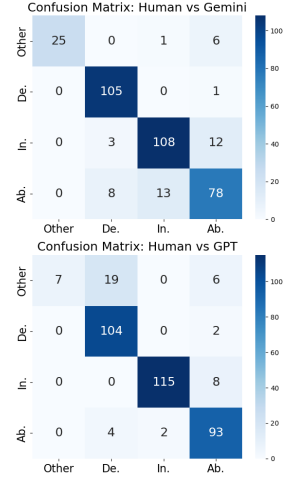


Figure 8: The confusion matrix between the aggregated human labels and the LLMs’ labels. The rows represents human predictions.

D LLM Judge Agreement

To verify the reliability of the automated judge, we prepare a validation dataset from a balanced subset of 360 responses, which are sampled across three model families and questions representing the three fundamental reasoning types. The annotators are asked to select the dominant reasoning type that directly contribute to the final answer. Two humans annotate the full validation dataset, then a third annotator reconcile their disagreement to form the gold standard. Both GEMINI-3-FLASH and GPT-5.1 demonstrate high alignment with human, achieving Cohen’s kappa (κ) scores of 0.83 and 0.84, respectively (Table 4). To further validate the LLM judge on steered model’s outputs, we conduct additional human evaluation on a subset of *OLMo3-7B-IT*’s results from α NLI (Fig. 5). The human evaluation results show that the compliance rate rises as the multiplier μ increases. Fig. 8 show the confusion matrix between the aggregated human labels and the LLMs’ labels. GEMINI-3-FLASH mostly fails on the abduction type, and often confuses it with the induction type. For GPT-5.1, the confusion between abduction and induction is less common, but it misclassified some deduction type into the “Other”. Alternatively, we also test prompting LLM judges with options of the reasoning types and their definitions, but that results in a lower agreement with the humans. Table 4 show the accuracy and κ of the LLM judges and human annotators.

Type	Prompt
Deductive	You are a logician tasked with performing deductive reasoning. You are given a general rule and specific observations. Your task is to apply the general rule to the observations to derive a logically certain conclusion. Provide a detailed reasoning process leading to your conclusion. [Question, Context, and Options...] Please enclose the answer in <answer></answer>.
Inductive	You are a logician tasked with performing inductive reasoning. You are given a set of observations. Your task is to infer the most probable general rule that explains all observations. Then, use the inferred rule to make predictions about new observations. Provide a detailed reasoning process leading to your conclusion. [Question, Context, and Options...] Please enclose the answer in <answer></answer>.
Abductive	You are a logician tasked with performing abductive reasoning. You are given a general rule and an observation. Your task is to generate the most probable and simplest hypothesis that, if true, would logically explain all the observations provided. Provide a detailed reasoning process leading to your conclusion. [Question, Context, and Options...] Please enclose the answer in <answer></answer>.

Table 3: Prompts for controlling LLMs’ reasoning.

Model	Accuracy	Cohen’s κ
GPT-5.1	.89	.84
Gemini-3-flash	.88	.83

Table 4: Agreement between the LLM judges and the human annotators.

μ	Compliance Rate
-1.0	24.2
0	81.8
1.0	87.9

Table 5: Human evaluation of the compliance rate with respect to the steering vector multiplier μ .

E Probing Results

Probing Instruction Encoding and Reasoning Compliance. Fig. 9 shows our probe analysis across the three open model families. A consistent pattern emerges from the results, the instructed reasoning type is almost perfectly decodable from the residual stream in all families, including on non-compliant cases. This suggests that failures of compliance are not primarily caused by a failure to represent the instruction. Instead, the models appear to encode the prompted reasoning paradigm early and robustly, but may still follow a competing internal trajectory when the instruction conflicts with the question’s logical structure. This pattern is strongest in LLaMA, where the gap between instruction decodability and judge-inferred reasoning is largest; OLMO shows the same trend with a smaller gap; and Qwen is the most aligned overall, with both high instruction decodability and the strongest judge-side signal. Taken together, these results support our main finding that the reasoning-instruction conflict is visible at the level of model representations. Models know what reasoning type they were asked to use, yet this does not guarantee compliance. On the judge side, the weaker proba-

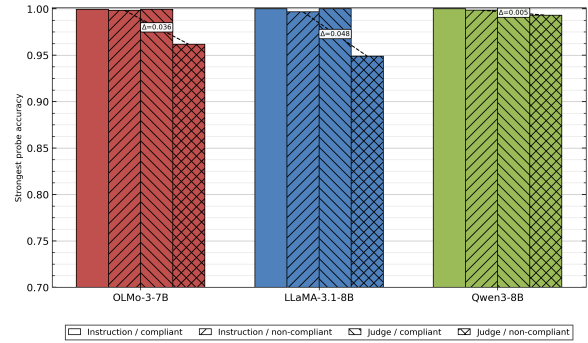


Figure 9: This figure shows the strongest probe accuracy for instructed reasoning type and judge-inferred reasoning type, split by compliant and non-compliant cases, across model families. A larger gap between the non-compliant instructed-reasoning bar and the non-compliant judge-reasoning bar indicates that the model represented the instruction clearly, but did not fully realize it in its final reasoning behavior.

bility of judge-inferred reasoning on non-compliant cases suggests that textual reasoning type is a noisier proxy for internal computation than the explicit instruction signal. Fig. 10 shows the full layer-wise accuracy results.

F Steering Results

Figure 11 show the steering results of LLaMA3.1-8B-IT on α -NLI. The compliance rates are moderately increased with higher μ values.

G Complete Accuracy Results

Table 6 shows the numeric accuracies of the judge-inferred $S \cap C$, $S \cap \neg C$, and $\neg S \cap C$ categories in the four datasets, which is corresponding to Figure 4. Table 7 shows the accuracy results grouped by the instructed reasoning type t' .

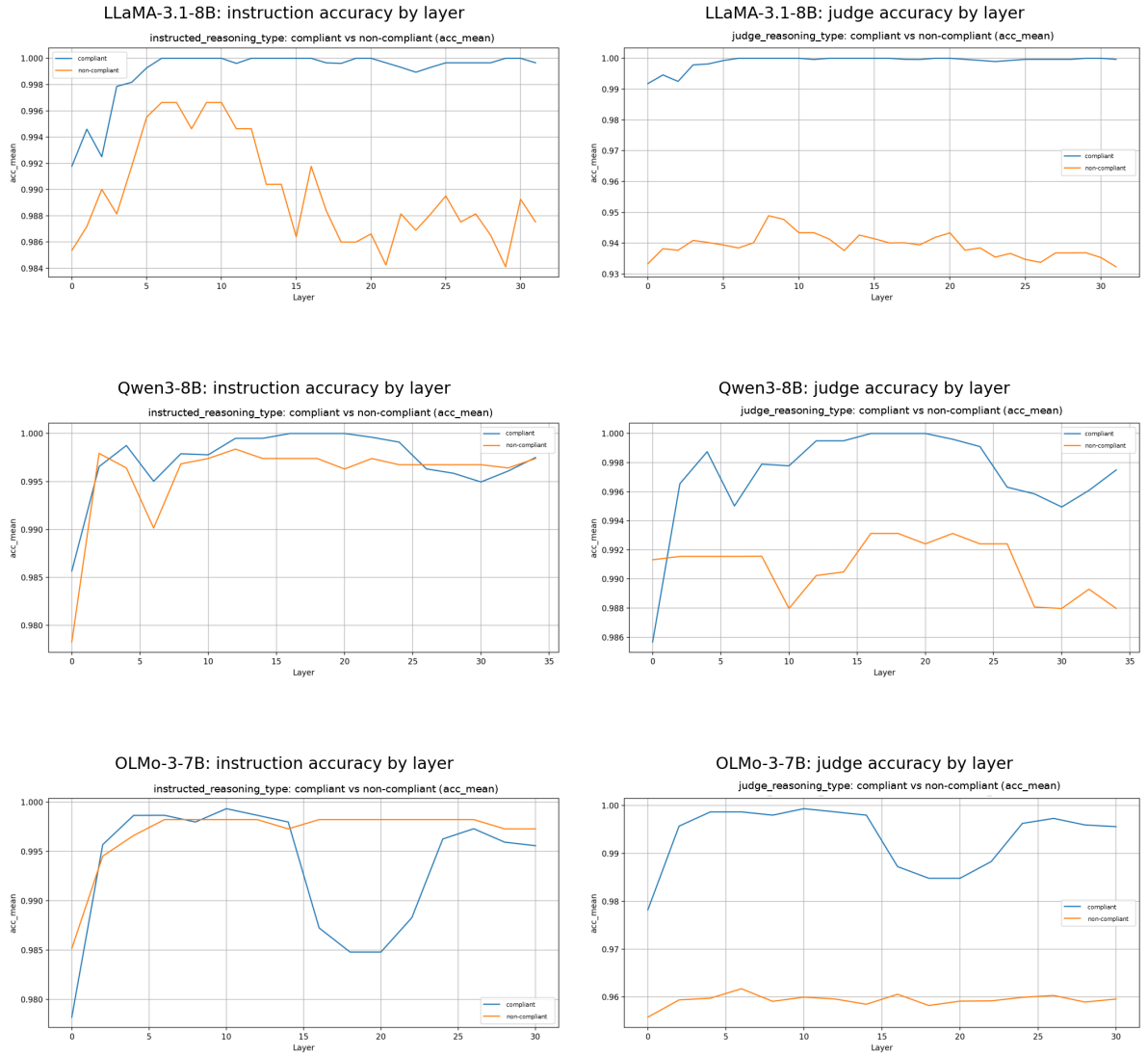


Figure 10: The layer-wise accuracy for each family of model and their instructed and judge-inferred reasoning types.

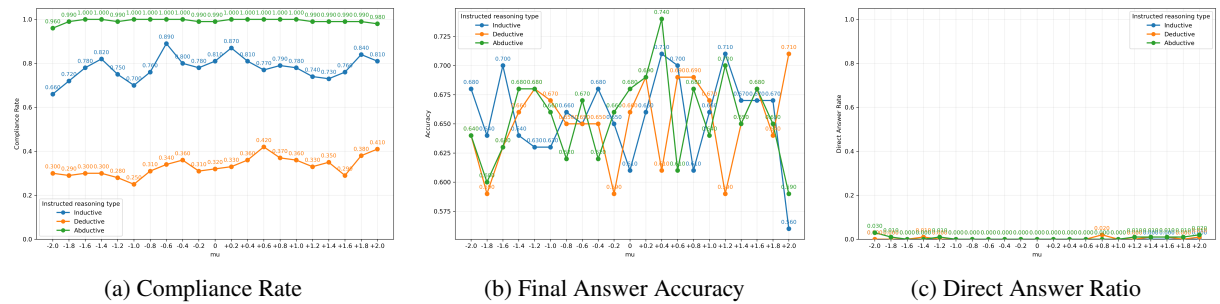


Figure 11: The impact of multiplier μ across reasoning types on α -NLI when steering the layer 14-17 of LLAMA3.1-8B-IT.

Data	LLM	S and C	S but not C	C but not S
FOLIO	GPT-5.1	76.81	71.67	n.a.
	GEMINI3-FLASH	88.50	87.37	42.11
	OLMO3.1-32B-IT	70.58	68.95	39.40
	LLAMA3.3-70B-IT	68.67	70.52	45.00
	QWEN3-32B	84.27	81.09	53.06
	OLMO3-7B-IT	49.26	54.62	41.82
	LLAMA3.1-8B-IT	58.60	61.58	46.12
SPR	QWEN3-8B	81.14	79.66	62.32
	GPT-5.1	67.79	70.56	50.00
	GEMINI3-FLASH	89.23	90.09	90.16
	OLMO3.1-32B-IT	69.32	69.32	46.88
	LLAMA3.3-70B-IT	66.44	70.42	52.63
	QWEN3-32B	64.36	66.28	41.72
	OLMO3-7B-IT	69.13	70.84	61.81
α_{NLI}	LLAMA3.1-8B-IT	38.11	45.24	23.35
	QWEN3-8B	57.88	69.41	37.65
	GPT-5.1	90.00	90.62	80.00
	GEMINI3-FLASH	85.57	86.67	89.68
	OLMO3.1-32B-IT	48.82	30.48	34.21
	LLAMA3.3-70B-IT	83.00	83.01	77.57
	QWEN3-32B	79.93	80.95	79.61
RECV	OLMO3-7B-IT	61.11	32.14	32.12
	LLAMA3.1-8B-IT	62.42	63.03	64.52
	QWEN3-8B	68.90	70.61	61.85
	GPT-5.1	78.75	80.66	81.18
	GEMINI3-FLASH	82.42	88.54	85.63
	OLMO3.1-32B-IT	77.78	74.42	69.39
	LLAMA3.3-70B-IT	80.79	83.41	84.83
	QWEN3-32B	77.78	80.84	80.87
	OLMO3-7B-IT	75.00	55.56	70.59
	LLAMA3.1-8B-IT	83.06	83.04	82.63
	QWEN3-8B	72.84	80.94	75.58

Table 6: The average accuracies of the sensible and compliant reasoning.

Data	LLM	DIRECT	DEDUCTIVE	INDUCTIVE	ABDUCTIVE
FOLIO	GPT-5.1	70.61 _{0.8}	76.68 _{0.6}	70.44 _{1.1}	71.10 _{1.4}
	GEMINI3-FLASH	88.01 _{0.7}	88.51 _{1.3}	79.47 _{1.2}	83.25 _{1.3}
	OLMO3.1-32B-IT	63.05 _{0.0}	76.85 _{1.1}	62.89 _{0.2}	65.35 _{0.8}
	LLAMA3.3-70B-IT	59.28 _{0.2}	68.97 _{1.1}	67.82 _{1.3}	69.13 _{0.6}
	QWEN3-32B	71.43 _{0.4}	85.55 _{0.5}	83.42 _{1.9}	81.94 _{1.2}
	OLMO3-7B-IT	54.35 _{1.8}	53.69 _{0.4}	52.96 _{0.7}	50.08 _{1.6}
	LLAMA3.1-8B-IT	53.86 _{1.7}	57.96 _{1.6}	57.47 _{3.3}	54.84 _{0.2}
SPR	QWEN3-8B	60.43 _{1.5}	80.95 _{1.2}	79.47 _{0.6}	77.83 _{1.2}
	GPT-5.1	68.67 _{0.5}	66.33 _{3.7}	67.33 _{3.7}	70.33 _{0.5}
	GEMINI3-FLASH	89.33 _{0.6}	89.67 _{0.5}	89.33 _{1.9}	90.33 _{0.4}
	OLMO3.1-32B-IT	43.67 _{0.5}	81.00 _{1.6}	57.00 _{0.8}	55.33 _{0.9}
	LLAMA3.3-70B-IT	46.67 _{0.5}	67.33 _{3.1}	65.67 _{4.2}	67.00 _{0.8}
	QWEN3-32B	60.67 _{0.5}	59.00 _{0.8}	59.33 _{2.9}	61.67 _{2.9}
	OLMO3-7B-IT	41.00 _{0.8}	70.67 _{0.9}	69.00 _{3.7}	68.00 _{1.4}
α_{NLI}	LLAMA3.1-8B-IT	34.67 _{2.5}	38.00 _{2.8}	37.00 _{1.6}	36.00 _{2.2}
	QWEN3-8B	40.00 _{0.0}	60.67 _{0.4}	58.00 _{2.4}	61.00 _{2.9}
	GPT-5.1	85.00 _{1.6}	89.00 _{0.8}	90.00 _{0.8}	90.00 _{0.8}
	GEMINI3-FLASH	90.33 _{1.5}	88.67 _{0.8}	88.33 _{0.9}	90.33 _{1.3}
	OLMO3.1-32B-IT	75.67 _{0.5}	33.33 _{0.9}	44.67 _{0.5}	59.33 _{0.5}
	LLAMA3.3-70B-IT	83.00 _{0.0}	83.33 _{0.5}	81.33 _{0.9}	83.00 _{0.8}
	QWEN3-32B	76.00 _{0.0}	83.33 _{1.2}	79.67 _{0.9}	81.00 _{0.8}
RECV	OLMO3-7B-IT	60.00 _{0.0}	25.33 _{2.5}	48.00 _{0.8}	57.33 _{3.1}
	LLAMA3.1-8B-IT	64.00 _{0.8}	67.00 _{1.6}	64.00 _{3.6}	62.67 _{2.5}
	QWEN3-8B	69.00 _{0.8}	67.33 _{2.5}	66.33 _{3.3}	69.00 _{0.8}
	GPT-5.1	83.16 _{1.3}	77.10 _{1.3}	83.84 _{0.8}	83.84 _{2.5}
	GEMINI3-FLASH	81.48 _{1.1}	80.47 _{0.8}	86.53 _{0.6}	86.53 _{1.9}
	OLMO3.1-32B-IT	72.39 _{0.5}	79.12 _{2.5}	80.47 _{1.0}	78.45 _{0.5}
	LLAMA3.3-70B-IT	79.80 _{0.0}	80.81 _{1.4}	83.84 _{0.8}	81.14 _{1.7}
	QWEN3-32B	82.49 _{0.5}	75.76 _{0.8}	78.11 _{0.5}	84.18 _{1.3}
	OLMO3-7B-IT	69.36 _{1.7}	67.00 _{2.1}	65.66 _{1.6}	68.01 _{2.5}
	LLAMA3.1-8B-IT	78.45 _{0.5}	80.47 _{3.3}	83.84 _{0.0}	83.50 _{1.0}
	QWEN3-8B	76.09 _{0.5}	72.05 _{1.9}	74.75 _{1.6}	76.09 _{1.9}

Table 7: The average accuracies of the final answer and their standard deviations with respect to the instructed reasoning types.