Do Cognitively Interpretable Reasoning Traces Improve LLM Performance?

Siddhant Bhambri[*]
SCAI, Arizona State University,
Tempe, US
siddhantbhambri@asu.edu

Upasana Biswas[*]
SCAI, Arizona State University,
Tempe, US
ubiswas2@asu.edu

Subbarao Kambhampati
SCAI, Arizona State University,
Tempe, US
rao@asu.edu

Abstract

Recent progress in reasoning-oriented Large Language Models (LLMs) has been driven by introducing Chain-of-Thought (CoT) traces, where models generate intermediate reasoning traces before producing an answer. These traces, as in DeepSeek R1, are not only used to guide inference but also serve as supervision signals for distillation into smaller models. A common but often implicit assumption is that CoT traces should be semantically meaningful and interpretable to the end user. While recent research questions the need for semantic nature of these traces, in this paper, we ask: "*Must CoT reasoning traces be interpretable to enhance LLM task performance?*" We investigate this question in the Open Book Question-Answering domain by supervised fine-tuning LLaMA and Qwen models on four types of reasoning traces: (1) DeepSeek R1 traces, (2) LLM-generated summaries of R1 traces, (3) LLM-generated post-hoc explanations of R1 traces, and (4) algorithmically generated verifiably correct traces. To quantify the trade-off between interpretability and performance, we further conduct a human-subject study with 100 participants rating the interpretability of each trace type. Our results reveal a striking mismatch: while fine-tuning on R1 traces yields the strongest performance, participants judged these traces to be the least interpretable. These findings suggest that it is useful to decouple intermediate tokens from end user interpretability.

1 Introduction

Reasoning with intermediate Chain-of-Thought (CoT)-style traces has become one of the defining strategies for improving the performance of Large Language Models over a diverse range of problems, as popularized by DeepSeek R1 [7]. While models such as DeepSeek R1 often generate excessively verbose responses [10], the R1-generated reasoning traces have been utilized as a learning signal for Supervised Fine-Tuning (SFT) smaller models to boost their performance [15, 19, 23].

A common but often implicit assumption behind these CoT traces is that they should be semantically meaningful and interpretable to humans. Although training with these traces is done primarily to improve LLM performance on a given task, these traces need not be semantically correct or

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: LAW 2025: Bridging Language, Agent, and World Models.

interpretable to optimize this objective. Moreover, since these reasoning traces are exposed to the end user, they can possibly exacerbate issues like user distrust, misinformation, errors, and perpetuated biases [6]. This distinction has also been brought to light by the recent GPT-OSS models that generate a CoT trace, a summary, and the final answer where the summary is shown for the humans and not the CoT trace [17]. There has been recent work that has challenged the first assumption behind these traces to be semantically meaningful by showing that both transformers and pre-trained LLMs can perform better when trained (or fine-tuned) with semantically incorrect traces paired with correct final solutions [3, 20]. In this work, we aim to specifically want to answer - "*Must CoT reasoning traces be interpretable to enhance LLM task performance?*".

We specifically look at the Open Book Question Answering domain and utilize the CoTemp QA benchmark [21] which consists of problems comprising a set of facts that can be utilized to answer the respective question. We conduct Supervised Fine-Tuning (SFT) experiments on Llama-3.2-1B-Instruct, Llama-3.1-8B, and Qwen3-1.7B, and Qwen3-8B chat models using four different variations of reasoning traces paired with correct final solutions. We consider (1) DeepSeek R1 traces, (2) LLM (GPT-4o-mini)-generated summaries of R1 traces, (3) LLM (GPT-4o-mini)-generated post-hoc explanations of R1 traces, and (4) algorithmically generated verifiably correct traces (as introduced in [3] for Open Book QA benchmarks). We then compare the final solution accuracy across all fine-tuned models.

To objectively compare the interpretability across each of these trace types, we further conduct a human-subject study with 100 participants. Five different sets of 25 participants were hired on Prolific and asked to rate each of the reasoning trace types on a Likert Scale measuring interpretability via attributes such as reasoning trace predictability, comprehensibility, and faithfulness [5, 9]. While fine-tuning on R1 traces shows the strongest performance on three out of the four LLMs, our striking results reveal that users find R1 traces to be the least interpretable across all tested attributes when compared with the other trace types. R1 traces scored the lowest among all variations of reasoning traces, averaging 3.396 among all interpretability attributes. These findings highlight that cognitive interpretability of reasoning traces can in fact be an albatross from the perspective of LLM's task performance.

2 Related Work

Large Language Models have significantly benefited from training with CoT traces coupled with final solutions for a variety of problems. There have been studies that have looked at and argued for making these CoT traces more interpretable, aka improve their faithfulness for the end user, as they are believed to serve as the LLM's explanations to generate the final solution [1, 22, 13, 24, 18, 14, 12, 26]. On the other hand, there has also been work showcasing why these traces are not explainable to the end user [2]. Both sides of this argument stem from the assumption that these traces are indeed meant to be useful for the end user and not just for the LLM to improve its final solution performance over a certain task. We specifically challenge this assumption in an effort to show the disconnect between the use of CoT traces for the LLM (as a training signal in SFT) and the use of CoT traces for the end user (as an interpretable reason behind the model's final solution). Interpretability has been studied along multiple attributes, including predictability, comprehensibility, interpretability, and faithfulness [9, 5]. These dimensions have been used to evaluate post hoc XAI methods and to develop a rigorous framework for interpretable machine learning.

3 Measuring LLM Performance via SFT w/ different Reasoning Traces

Dataset and Metrics: CoTemp QA [21] consists of English co-temporal questions which involve identifying the type of temporal relation posed in the problem, followed by inferring which fact in the given passage of text satisfies the temporal relation with the question. For all our SFT experiments, we look at the final solution accuracy across all fine-tuned models using the four trace types.

Reasoning Trace generation: We first consider (1) DeepSeek R1 traces where we prompt the R1 model on the CoTemp QA training dataset and collect the model responses for our SFT experiments where it got the correct final answer. Utilizing this filtered training dataset, we prompt GPT-4o-mini to generate both (2) summaries and (3) post-hoc explanations of these R1 traces. Since R1 traces can often be verbose, we posit that their summary as well as a post-hoc explanation can likely be

more interpretable to the end user. As a control study, we utilize the generated SFT trace datasets from [3] which consist of the (4) semantically correct verifiable reasoning traces. These traces have been algorithmically generated by extracting the relevant fact/s from the provided passage.

Results: We highlight the key SFT results in Figure 1. A common observation seen across three out of the four models (except in Qwen3-8B) is that SFT with R1 traces leads to the highest final solution accuracy over SFT with any other trace type with the largest performance boost seen for Llama-3.2-1B-Instruct model. Furthermore, among all the four models, we note that SFT with the algorithmically-generated semantically correct traces and SFT with the adversarially-generated incorrect traces perform the worst also in comparison to SFT with summaries and explanations of R1 traces. Keeping these results in consideration, we conduct a user study to test if the R1 traces that led to the best performing SFT models are interpretable to humans.

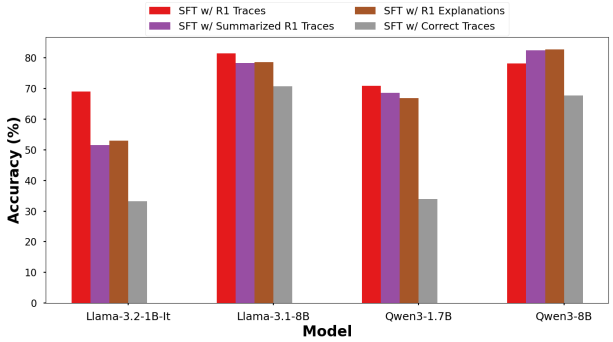


Figure 1: Final solution performance on CoTemp QA test dataset after SFT with different trace types on Llama and Qwen models.

4 Measuring Trace Interpretability via Human-Subject Studies

We conducted four separate user studies to evaluate the interpretability of the four types of reasoning traces. In each study, participants were shown only one type of trace: (1) DeepSeek R1 traces, (2) summarized R1 traces, (3) post-hoc explanations of R1 traces, or (4) verifiably correct reasoning races [3]. Our hypotheses are tested in a between-subjects design, comparing participant responses across these four groups. The specific hypotheses we test are:

- **H1:** Reasoning traces that improve task accuracy will not lead to higher interpretability for the user.
- **H2:** Reasoning traces that improve task accuracy will be associated with higher cognitive workload for the user, as measured by increased mental demand, effort, and frustration.

Experimental Setup: For each trace type, we recruited 25 users (100 in total) on Prolific. Each participant viewed five Q/A examples (fixed across all studies), consisting of the input question, the predicted answer, and the reasoning trace. After each example, participants rated the reasoning trace on a 5-point Likert scale on the following properties as suggested by [5, 9]: predictability, comprehensibility, interpretability, and faithfulness to context (alignment with given facts). To capture the cognitive workload involved in processing and evaluating the traces, we used the NASA–TLX assessment [8], focusing on the dimensions of mental demand, effort, and frustration.
Main Findings: From Table 1, we observe that participants rated algorithmically generated correct reasoning traces from [3] as the most interpretable across all dimensions—predictability, comprehensibility, interpretability, and faithfulness—consistently scoring higher medians than all other trace types. In contrast, R1 traces received the lowest interpretability ratings across every dimension. Summarized R1 traces and R1-trace explanations received intermediate ratings, indicating that compact or post-hoc representations improve human comprehension compared to raw R1 traces. In terms of cognitive workload, R1 traces imposed higher mental demand, effort, and frustration compared to other kinds of traces. Correct traces were associated with relatively lower cognitive workload, indicating that users found them easiest to follow and comprehend.
Statistical Analysis: We conducted pairwise Mann–Whitney U tests [16] at a significance level of $\alpha = 0.05$ with Bonferroni correction applied for multiple comparisons. Null hypotheses derived from our study hypotheses were defined as follows: NH1 (Interpretability): There is no difference in interpretability ratings between R1 traces and algorithmically-generated correct reasoning traces. NH2 (Cognitive Workload): There is no difference in cognitive workload ratings between R1 traces

and algorithmically-generated correct reasoning traces.
There was a significant difference in interpretability measured between these two trace types, across all measured dimensions (predictability: U = 176.5 ,$p = .00022 < 0.05$; comprehensibility: U = 175, $p = .00019 < 0.05$; interpretability: U = 161, $p = .00014 < 0.05$; faithfulness: U = 178.5, $p = .00015 < 0.05$). Further analysis also shows that there was a significant difference between cognitive workload of users between the two trace types, across all measured dimensions (mental demand: U = 194, $p = .00036 < 0.05$; effort: U = 176, $p = .00013 < 0.05$; frustration: U = 176.5, $p = .01287 < 0.05$). Thus, we can reject NH1 and NH2, validating

Dimension	Question	R1 Traces	Summarized R1 Traces	R1 Explanations	Correct Traces
Predictability	I could anticipate the next steps or conclusions based on earlier parts of the reasoning ↑	**3.48**	4.45	4.29	4.82
Comprehensibility	I understood the reasoning followed by the model ↑	**3.46**	4.55	4.27	4.56
	I could follow each step in the reasoning without confusion ↑	**3.46**	4.54	4.28	4.84
Interpretability	The reasoning helped me understand why the model acted or concluded the way it did ↑	**3.31**	4.53	4.29	4.86
Faithfulness	There were no major gaps or missing reasoning steps in the reasoning ↑	**3.33**	4.54	4.26	4.72
	The reasoning is consistent with the facts or evidence provided in the context ↑	**3.34**	4.24	4.29	4.84
Mental Demand	How mentally demanding was the task? ↓	**4.65**	2.87	2.92	2.31
Effort	How hard did you have to work to accomplish your level of performance? ↓	**4.54**	2.39	2.17	2.86
Frustration	How frustrated, stressed, and annoyed were you? ↓	**4.58**	2.04	2.42	2.42

Table 1: Median participant ratings of reasoning traces across dimensions of interpretability and cognitive workload. Arrows indicate the desired direction of scores: ↑ higher ratings are better for interpretability measures, ↓ lower ratings are better for cognitive workload measures.

5 Discussion

Our findings reveal a major disconnect between the utility of reasoning traces for improving LLM performance and their cognitive interpretability for humans. SFT with R1 traces led to higher accuracy, yet these traces were rated the lowest across all dimensions of interpretability—predictability, comprehensibility, and faithfulness—and were associated with the highest mental demand, effort, and frustration. Summaries and post-hoc explanations of R1 traces further validate this point: although they yielded lower accuracy than R1 traces, they are easier for users to predict, understand, and perceive as faithful. By contrast, algorithmically generated correct traces were judged as most interpretable and least mentally demanding, but yielded the weakest improvements in model accuracy. These results clearly highlight that verbose traces like R1 provide rich training signals for models, but are poorly aligned with the interpretability expectations of the end user. Furthermore, the reasoning traces which benefit the LLMs the most need not have semantic structure, underscoring a fundamental disconnect between what serves as a good training signal and what supports human understanding.

6 Conclusion

In this work, we studied the relationship between the use reasoning traces for improving model performance and their interpretability for end users. Through SFT experiments on four LLMs, we observed that R1 traces yielded the highest accuracy on the CoTemp QA benchmark over other trace types. In contrast, our human-subject study revealed that these same R1 traces were rated lowest across all interpretability dimensions and imposed the greatest cognitive workload. These findings demonstrate that reasoning traces which help models do not necessarily carry semantics that humans find interpretable. This decoupling broadly has two key takeaways for future works - (1) CoT-style traces should only be utilized for optimizing model performance and not end-user interpretability, and (2) independent efforts should be carried out for generating explanations behind the model's answer tailored for the end-user.

Acknowledgment

This research is supported in part by grants from ONR (N00014-25-1-2301 and N00014-23-1-2409), DARPA (HR00112520016), DoD RAI (via CMU subcontract 25-00306-SUB-000), an Amazon Research Award, and a generous gift from Qualcomm.

References

[1] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

[2] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v2, 2025.

[3] Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation. *arXiv preprint arXiv:2505.13792*, 2025.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[5] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[9] Anahid N. Jalali, Bernhard Haslhofer, Simone Kriglstein, and Andreas Rauber. Predictability and comprehensibility in post-hoc xai methods: A user-centered analysis. *ArXiv*, abs/2309.11987, 2023.

[10] Subbarao Kambhampati, Kaya Stechly, and Karthik Valmeekam. (how) do reasoning models reason? *Annals of the New York Academy of Sciences*, 2025.

[11] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[12] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[13] Jiachun Li, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Towards better chain-of-thought: A reflection on effectiveness and faithfulness. *arXiv preprint arXiv:2405.18915*, 2024.

[14] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

[15] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

[16] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[17] OpenAI. Introducing gpt-oss. `https://openai.com/index/introducing-gpt-oss/`, 2025. Accessed: 2025-08-21.

[18] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2402.13950*, 2024.

[19] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[20] Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775*, 2025.

[21] Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*, 2024.

[22] Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.

[23] Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 251–260, 2025.

[24] Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. *arXiv preprint arXiv:2502.14829*, 2025.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[26] Wei Jie Yeo, Ranjan Satapathy, Rick Siow Mong Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? *arXiv preprint arXiv:2402.11863*, 2024.

A Additional Experiment Details

A.1 Dataset

CoTemp QA: The dataset is categorized into four temporal relation types, namely - 'equal', 'overlap', 'during' and 'mix', and requires around one or two facts for answering the question. For our experiments, we utilize 3,798 train and 950 test samples to construct the SFT datasets. The train/test splits for each category are shown in Table 2.

Table 2: Train and Test data distribution for CoTemp QA dataset used in our SFT experiments.

Category	Train/Test Samples
equal	349 / 87
overlap	522 / 131
during	2477 / 619
mix	450 / 113

A.2 Implementation Details and Hyper-parameters

Models were fine-tuned using the Hugging Face library [25] on a single 80GB NVIDIA Tesla A100 GPU for 3 epochs (effective batch size 16, max sequence length 1024). We employed PEFT QLoRA [4] (rank 16, alpha 32) with a learning rate of 2e-4 (8-bit AdamW, cosine scheduler, 0.1 warm-up). Prompt experiments utilized vLLM [11]. We will release the code and datasets used for our experiments on acceptance.

A.3 Prompts

R1 Trace Summarization Prompt

Summarize the following trace in a very concise and clear manner, highlighting key events and outcomes in less than 100 words:
...
{R1 trace}
...
Summary:

R1 Trace Explanation Prompt

{Problem}
...
{R1 trace}
...
{R1 answer}
You have answered the question correctly. Please provide a detailed explanation of the reasoning behind your answer. The explanation should be clear, concise, and easy to understand.
...
Explanation:

B User Study

To evaluate the interpretability of reasoning traces generated by reasoning models, we conducted a set of structured user studies. Each participant was given a compensation of $12\$/hr$. The IRB protocol details will be released on acceptance. Each study followed the same sequence of steps, designed to ensure consistency across participants for each study. Below we outline the main components of the study design.

B.1 Human Participant Demographics

We conducted four user studies with participants recruited through Prolific (all located in the United States). In general, the participant populations in all four studies were demographically similar, with no major differences in the age or education distribution, suggesting that the results in the studies are comparable and not driven by differences in the composition of the participants.

Education: Participants spanned a range of educational backgrounds. Across all studies, the majority held an *Undergraduate Degree* (roughly 45–55% in each study), followed by *Master's Degrees* (20–30%), and a smaller proportion with *PhDs or equivalent doctoral-level degrees* (10–15%). A minority of participants reported *High School*, *Associate's Degree*, or *Some College* as their highest level of education (<10% each). These proportions were consistent across the four studies.

Age: The participants were distributed over a wide age range, with the largest groups being *35–50 years old* (approximately 35–40%) and *51+ years old* (30–35%). Younger age groups were represented to a lesser extent: *26–34 years old* (20–25%) and *18–25 years old* (5–10%). Again, these proportions were stable across studies.

B.2 Consent and Statement

Each participant began the study by reviewing and agreeing to a consent statement. The statement explained the goals of the study, what participants would be asked to do, and how their data would be handled.

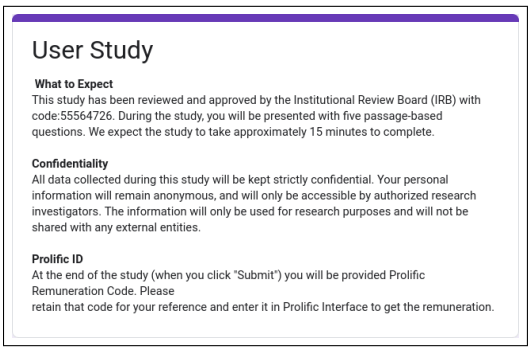


Figure 2: Consent statement shown to participants before starting the study.

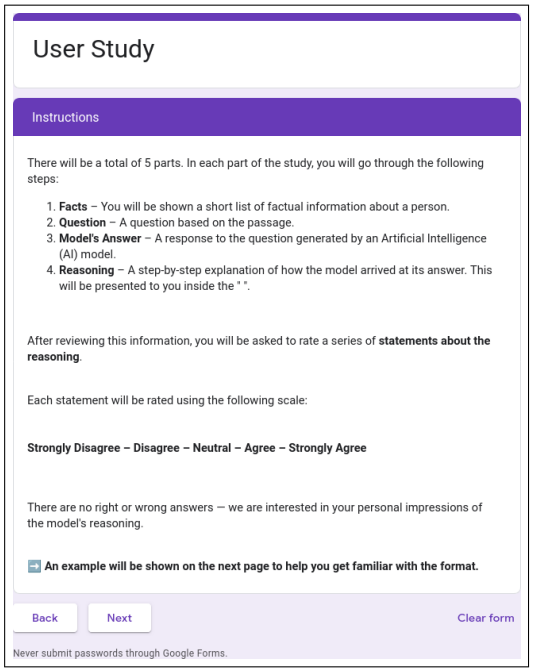


Figure 3: Instructions shown to participants before starting the study.

Figure 4: Consent statement (left) and instructions (right) shown to participants.

B.3 Instructions

Participants were provided with detailed instructions describing the study structure. Each of the five parts of the study followed the same format:

1. **Facts:** A short list of factual statements about a person.

2. **Question:** A query based on the passage.
3. **Model's Answer:** The response generated by the AI model.
4. **Reasoning:** A step-by-step explanation of how the model arrived at its answer.

After reviewing this information, participants rated statements about the reasoning on a 5-point Likert scale (Strongly Disagree–Strongly Agree).

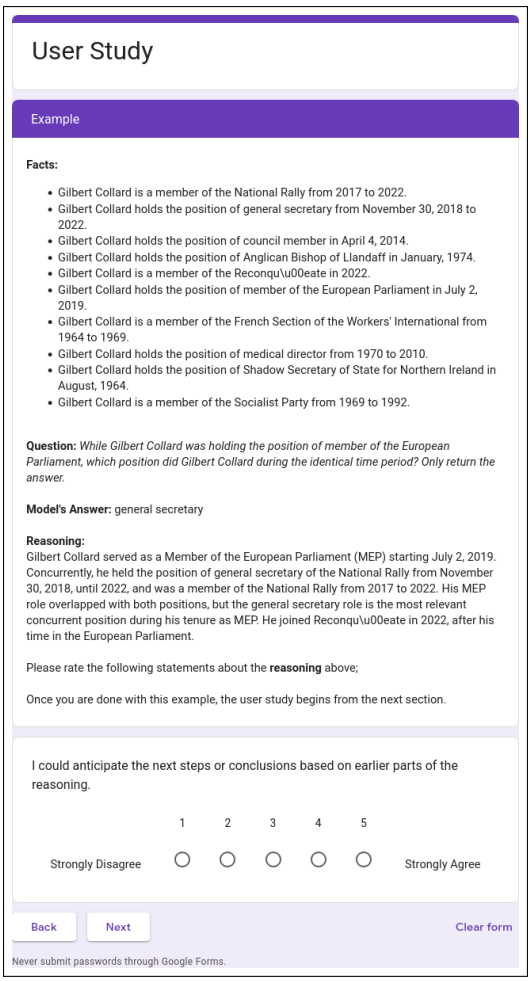


Figure 5: Example shown to participants.

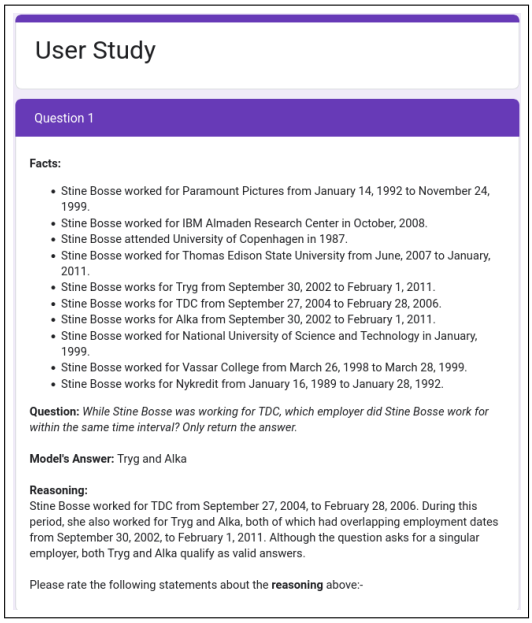


Figure 6: Task shown to participants.

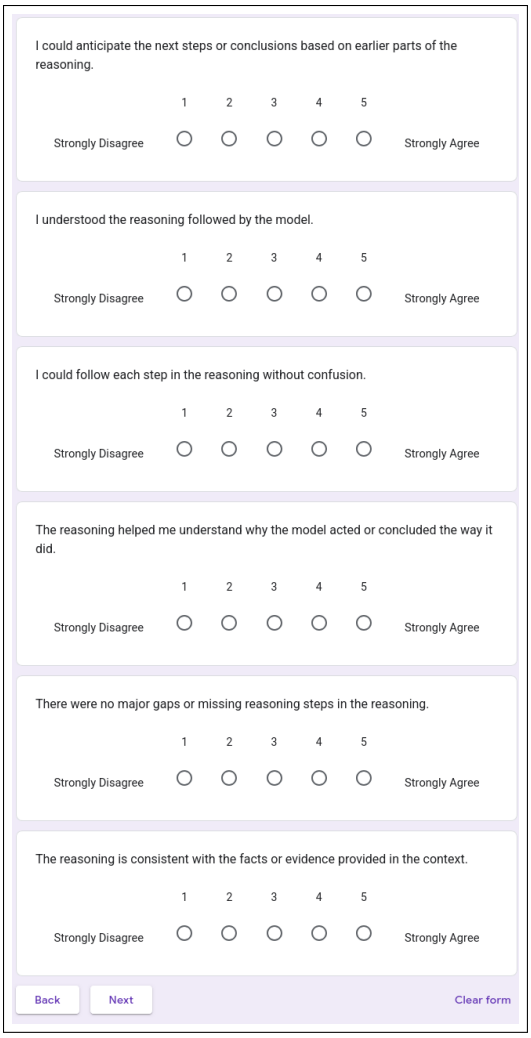
Figure 7: Example (left) and task (right) shown to participants.

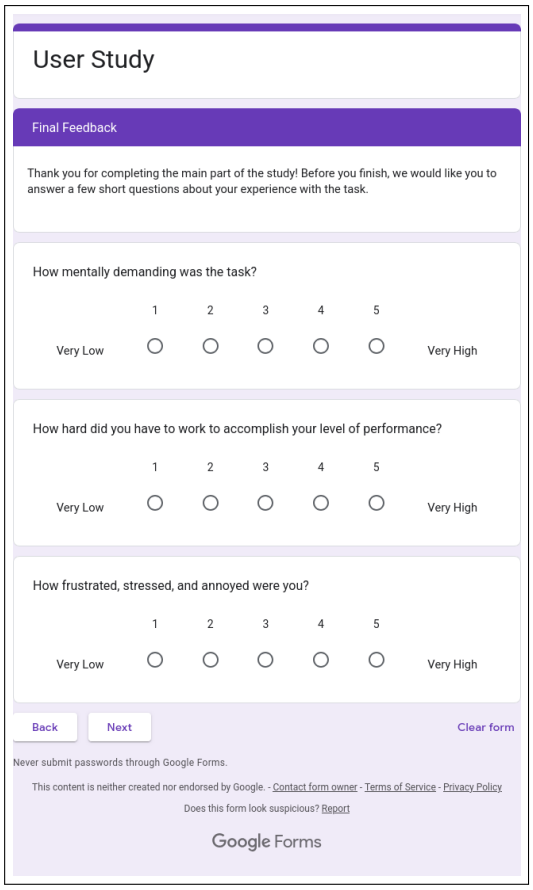
B.4 Q/A Task

Before beginning the main task, participants reviewed an example question and answer with reasoning (see Fig. X). Participants then completed five Q/A tasks of the same form as the example. Each task included a passage, model answer, reasoning trace, and associated questionnaire.

B.5 Questionnaire

After each reasoning trace, participants filled out a brief questionnaire assessing dimensions of interpretability (see Fig. Y). At the end of the study, participants completed a feedback survey, with NASA-TLX questions to measure perceived workload.





B.6 Statistical Analysis Results

Dimension	R1 vs Correct			R1 vs Summarized R1			R1 vs Explanations		
	U	p	Sig.	U	p	Sig.	U	p	Sig.
H1: Interpretability									
Predictability	176.5	.00022	Yes	177.0	.00036	Yes	126.5	.0004	Yes
Comprehensibility	175	.00019	Yes	102.2	$< .00001$	Yes	126	.0006	Yes
Interpretability	161	.00014	Yes	74.5	$< .00001$	Yes	187	$< .00001$	Yes
Faithfulness	178.5	.00015	Yes	73.5	$< .00001$	Yes	115.5	$< .00001$	Yes
H2: Cognitive Workload									
Mental Demand	194	.00036	Yes	237.5	0.055	No	264	.21	No
Effort	176	.00013	Yes	169.5	.0016	Yes	104	$< .00001$	Yes
Frustration	102.5	.01287	Yes	164	0.0013	Yes	158	.00056	Yes

Table 3: Pairwise Mann–Whitney U test results across different trace types. Significance is determined at $\alpha = 0.05$ with Bonferroni correction.

C Limitations

In this work, we analyze the correlation between the final solution performance of LLMs when fine-tuned with different types of CoT-style traces and the interpretability of these traces for end users. Due to computational limitations, we restrict our scope on experiments with models up to 8 billion parameters. We also restrict our study on simple Open Book QA problems that can be easily understood and answered by lay users. Future works can scale our study to analyze the impact on final performance by fine-tuning larger parameter LLMs, and hire domain experts for conducting user studies on scientific benchmarks such as math or coding problems.