Interpretable Traces, Unexpected Outcomes: Investigating the Disconnect in Trace-Based Knowledge Distillation

Anonymous ACL submission

Abstract

Recent advances in reasoning-oriented Large Language Models (LLMs) have been driven by the introduction of 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 model inference but also serve as supervision signals for Knowledge Distillation (KD) to improve smaller models. A prevailing but under-examined implicit assumption is that these CoT traces are both semantically correct and interpretable for the end-users. While there are reasons to believe that these intermediate tokens help improve solution accuracy, in this work, we question their validity (semantic correctness) and interpretability to the end user. To isolate the effect of trace semantics, we design experiments in the Question Answering (QA) domain using a rule-based problem decomposition method. This enables us to create Supervised Fine-Tuning (SFT) datasets for LLMs where - each QA problem is paired with either verifiably correct or incorrect CoT traces, while always providing the correct final solution. Trace correctness is then evaluated by checking the accuracy of every sub-step in decomposed reasoning chains. To assess end-user trace end-user interpretability, we also fine-tune LLMs with three additional types of CoT traces: DeepSeek R1 traces, LLM-generated summaries of R1 traces, and LLM-generated post-hoc explanations of R1 traces. We further conduct a human-subject study with 100 participants asking them to rate the interpretability of each trace type on a standardized Likert scale. Our experiments reveal two key findings - (1) Correctness of CoT traces is not reliably correlated with the model's generation of correct final answers: correct traces led to correct solutions only for 28% test-set problems while incorrect traces don't necessarily degrade solution accuracy. (2) In end-user interpretability studies, fine-tuning on verbose DeepSeek R1 traces produced the best model performance but these traces were rated as least interpretable by users, scoring on average 3.39 for interpretability and 4.59 for cognitive load metrics on a 5-point Likert scale. In contrast, the decomposed traces that are judged significantly more interpretable don't lead to comparable solution accuracy. Together, these findings challenge the assumption in question suggesting that researchers and practitioners should decouple model supervision objectives from end-user-facing trace design.

1 Introduction

Reasoning with intermediate Chain-of-Thought (CoT)-style traces (step-by-step outputs that models produce prior to an answer) has become one of the defining strategies for improving the performance of Large Language Models (LLMs) over a diverse range of problems, as exemplified by approaches like DeepSeek R1 (Guo et al., 2025). While models such as DeepSeek R1 often produce extremely verbose unstructured responses even for simple problems (Kambhampati et al., 2025), these reasoning traces are utilized both as inference aids and supervision signals in Knowledge Distillation (KD) when Supervised Fine-Tuning (SFT) smaller LLMs for enhanced task performance (Magister et al., 2022; Shridhar et al., 2022; Tian et al., 2025).

A common but often implicit assumption behind these CoT traces is that they are semantically correct and interpretable for end-users (Guo et al., 2025). Training with these traces is done primarily to improve LLM performance on a given task, but training/fine-tuning objectives rarely require these traces to be semantically correct or interpretable. In this work, we challenge this assumption and ask: "*Must CoT reasoning traces be semantically correct and interpretable to end-user for enhancing LLM task performance?*"

To address this, we focus our experiments on the Question Answering (QA) domain, where end-users regularly interact with both intermediate

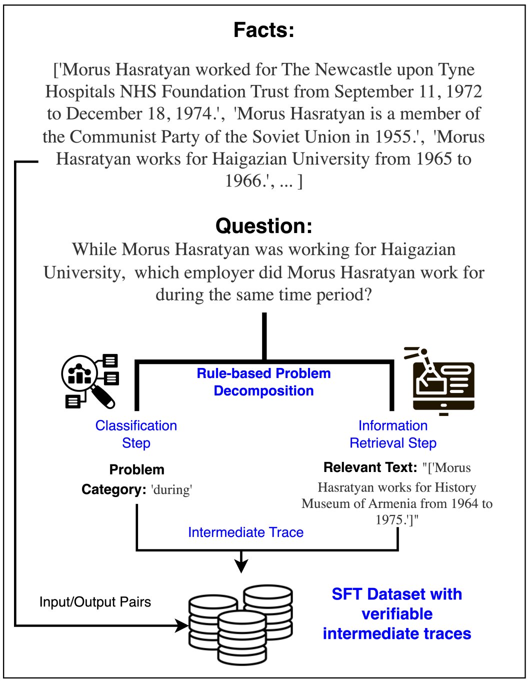


Figure 1: The construction of SFT dataset w/ verifiable intermediate traces using rule-based problem decomposition on an example from the CoTemp QA dataset.

traces and final outputs (Polemi et al., 2024) (e.g., ChatGPT (OpenAI, 2023), Perplexity (AI, 2023), Copilot (Microsoft, 2023), Gemini (Google, 2023)). Faithfulness of reasoning traces is especially critical in these interactive settings where unverifiable traces can lead to loss of trust in users, misinformation and errors in model outputs, and perpetuation of biases among other negative consequences (Guidotti et al., 2018). To assess the trade-offs between semantic correctness of the traces and LLM performance, we design an experimental setting where both final solutions and intermediate traces can be independently evaluated. Specifically, we employ a rule-based problem decomposition technique to break QA tasks into structured subproblems (McDonald and Emami, 2024; Xue et al., 2024). Next, we generate SFT datasets pairing questions with either verifiably correct or verifiably incorrect reasoning traces (while always including the correct answer). At inference, this allows us to verify the correctness of both the final solution and the intermediate traces generated by the distilled model.

To assess the trade-offs between end-user interpretability and LLM performance, we fine-tune models on different types of reasoning traces: DeepSeek R1 traces (verbose CoT outputs), LLM (GPT-4o-mini)-generated summaries of R1 traces (end-user facing summarizations), LLM (GPT-4o-mini)-generated post-hoc explanations (natural language explanations of R1 traces), and verifiably correct traces that we discussed above. In parallel, we conduct a human-subject study with 100 participants (hired on Prolific), split into four sets of 25. Each group was asked to judge the interpretability of the trace types using a Likert Scale measuring predictability, comprehensibility, and faithfulness attributes (Jalali et al., 2023; Doshi-Velez and Kim, 2017).

Our experiments reveal two key findings: (1) **Correctness of CoT traces** is not reliably correlated with LLMs producing correct final answers: correct traces led to correct solutions only for 28% test-set problems, while incorrect traces did not consistently degrade answer accuracy. (2) **End-user interpretability** of CoT traces is not reliably correlated with LLMs producing correct final answers: fine-tuning on verbose DeepSeek R1 traces led to the strongest task performance, yet users rated these traces as least interpretable, scoring on average 3.39 for interpretability and 4.59 for cognitive load metrics on a 5-point Likert scale. These results highlight that **semantic correctness and human interpretability of reasoning traces can in fact be an albatross from the perspective of LLM's task performance**, challenging assumptions in current LLM supervision practices.

The paper is organized as follows: §2 reviews prior work on Large Language & Reasoning Models, Knowledge Distillation, and CoT Trace Interpretability. §3 presents our problem setup, rule-based decomposition for Open-Book QA, and dataset construction for distilling LLMs with correct and incorrect intermediate traces. §4 describes the SFT experiments and human-subject studies, and §5 analyzes results and key insights. We conclude the work in §6. Supplementary materials include additional experiments, and all datasets and code will be released upon acceptance.

2 Related Work

2.1 Large Reasoning Models & CoT traces

Large Language Models (LLMs) have shown remarkable performance on a wide variety of natural language tasks in question answering, text generation, summarization, and translation, to name a few (Bubeck et al., 2023). Recent advances in post-training techniques have led to the rise of Large Reasoning Models (LRMs) such as DeepSeek R1

(Guo et al., 2025), Google Gemini 2.5 (Google, 2023), Microsoft Phi-4-reasoning (Abdin et al., 2025), etc. These reasoning models produce a set of intermediate tokens, commonly referred to as 'reasoning' traces, followed by the final solution. While LRMs have shown a significant improvement in final solution accuracy on reasoning tasks over standard LLMs (Guo et al., 2025; Abdin et al., 2025), their intermediate traces are subjective and verbose (Wei et al., 2022; Wang et al., 2022), making it hard to evaluate their trace validity and end-user interpretability (Kambhampati et al., 2025).

2.2 Knowledge Distillation

While Small Language Models (SLMs) offer a computationally efficient alternative to LLMs and LRMs, they are not robust to prompt augmentations (such as Chain-of-Thought) or steerable using in-context examples used in few-shot prompt settings (Shridhar et al., 2022; Stolfo et al., 2022). Knowledge Distillation is a well-studied approach used for fine-tuning these SLMs (student) via the outputs of a larger model (teacher) (Magister et al., 2022). With LRMs generating both an intermediate trace and the final solution, SLMs are also distilled to replicate this output (Shridhar et al., 2022; Tian et al., 2025). However, the lack of structured intermediate trace outputs makes the validity of the traces hard to evaluate (Zhou et al., 2022; Chen et al., 2022). This problem is exacerbated for end-user settings such as in Question Answering (QA) domains, where user interactions involve exposure to both intermediate traces and final outputs.

2.3 Interpretability of CoT Traces

Some recent works have argued for making these CoT traces more interpretable to the end-user, i.e., improve their faithfulness for the end user, as they are believed to serve as the LLM's explanation to generate the final solution (Arcuschin et al., 2025; Tanneru et al., 2024; Li et al., 2024; Tutek et al., 2025; Paul et al., 2024; Lyu et al., 2023; Lanham et al., 2023; Yeo et al., 2024). On the other hand, there has also been work showcasing why these traces are not explainable to the end user (Barez et al., 2025). Both sides of this argument stem from the assumption that these traces are indeed meant to be useful and interpretable 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 and 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).

3 Knowledge Distillation using Problem Decomposition

This section describes our rule-based problem decomposition method for breaking complex Open Book QA tasks into verifiable sub-problems (§3.1) and explains how we use it to generate structured intermediate traces for SLM distillation (§3.2).

3.1 Rule-based Problem Decomposition

In the context of Open Book QA, consider the example shown in Figure 1 which consists of a text passage (referred to as set of facts for our discussion) and a question involving temporal reasoning between the queried problem and the facts present in the provided text. Answering this reasoning question involves identifying the relevant fact from the text which satisfies the temporal relation asked in the problem. In this case, the queried fact refers to *"Morus Hasratyan works for Haigazian University from 1965 to 1966."* The temporal relation queried in the problem is 'during' and thus, the relevant fact that answers the query is *"Morus Hasratyan works for History Museum of Armenia from 1964 to 1975."* Hence, the final answer is *'History Museum of Armenia'*. From this example, we see that the complex Open Book QA problem can be decomposed into a 1) Classification step determining the type of question asked ('during' temporal relation in this case), and an 2) Information Retrieval (IR) step to determine the relevant part of text that can answer the query (the fact with the temporal overlap with the one in question). Therefore, we utilize these two steps to decompose the Open Book QA problems that allow us to construct structured intermediate traces for evaluation.

3.2 Intermediate Trace Generation for SFT

Given the outputs of the sub-problems obtained by decomposing the original query as shown in Figure 1, we generate the intermediate traces in an automated way which consists of the Classification step describing the type of the question posed in the query, and the IR step showing the relevant fact in the text that can help answer the query. We construct a dataset using these Input-Trace-Output tuples that can be utilized to SFT the Small Language Models. Note, that by constructing the intermediate trace using these two steps, we can then evaluate the accuracy of the intermediate traces generated by the distilled model at the time of inference. We will refer to this setting as **SFT w/ Correct Traces** for further discussion.

To critically understand the correlation between intermediate trace correctness and final solution accuracy for Knowledge Distillation methods, we also consider an alternative SFT setting where for every input problem, we choose an incorrect problem category and incorrect fact/s for constructing the intermediate trace. This allows us to construct a SFT dataset which also consists of Input-Trace-Output tuples but with incorrect traces and correct final outputs. We will refer to this setting as **SFT w/ Incorrect Traces**. We discuss the empirical setup for our experiments in the following section.

4 Experimental Setup

This section presents the Open Book QA datasets used in our experiments (§4.1), details our analysis of the link between trace semantic correctness and LLM performance (§4.2), and describes the SFT experiments and human study evaluating interpretability–performance trade-offs (§4.3). Implementation details are described in §4.4.

4.1 Datasets

CoTemp QA: CoTemp QA (Su et al., 2024) is an English dataset of co-temporal questions requiring identification of a temporal relation type and inference of the corresponding fact in a passage. It includes four relation types—equal, overlap, during, and mix—and typically needs one or two supporting facts per question. We use 3,798 training and 950 test samples for our SFT experiments.

Microsoft MARCO QA: The Microsoft MAchine Reading COmprehension (MARCO) dataset (Bajaj et al., 2016) is an English dataset of real user queries from Bing, each accompanied by long passages sourced from supporting URLs. It includes five query types—description, numeric, entity, location, and person—and typically requires one paragraph to answer. We use 5,000 training and 1,000 test samples in our experiments.

Facebook bAbI QA: The Facebook bAbI QA dataset (Weston et al., 2015) is an English benchmark for evaluating reading comprehension through QA tasks that test reasoning skills such as fact chaining and deduction. While the full dataset contains 20 categories, we use 11 in our experiments—single-supporting-fact, two-supporting-facts, two-arg-relations, counting, lists-sets, conjunction, time-reasoning, basic-deduction, basic-induction, positional-reasoning, and size-reasoning. Each question typically requires about three supporting facts. Our SFT dataset includes 3,773 training and 376 test samples, with detailed splits provided in §A.1.

4.2 Trace Correctness vs LLM Task Performance

For our experiments to evaluate the correlation between intermediate trace correctness and final solution accuracy, we utilize the Llama-3.2-1B-Instruct and the Qwen3-1.7B chat models. We adopt the following baselines for our evaluations:

Direct Prompting SLMs: We directly prompt the two SLMs to establish the baseline performance of these models across the three datasets without any additional fine-tuning.

SFT - Vanilla: Following the conventional fine-tuning technique, we also utilize the SFT baseline where we fine-tune the models using only Input-Output pairs and no intermediate traces. This allows us to evaluate the final solution performance for these models against the final solution performance obtained via directly prompting and via SFT with intermediate traces.

For examining the impact and correctness of intermediate traces, we run the following experiments:

SFT w/ Correct Traces: Using the intermediate traces constructed via problem decomposition (Section 3.2), we fine-tune the models using the Input-Trace-Output tuples for each of the three datasets.

SFT w/ Incorrect Traces: In this case, we construct incorrect intermediate traces as discussed in Section 3.2, but use the correct final solutions in the Input-Trace-Output tuples.

For our experiments on SFT w/ Correct Traces and SFT w/ Incorrect Traces, we additionally report Category Accuracy (signifying the Classification step performance in the intermediate trace), the IR step Accuracy (signifying the IR step performance in the intermediate trace), and average trace length (# of words in the intermediate trace) on the test datasets computed at inference time.

4.3 Interpretability of Traces vs Final Solution Accuracy

To assess the correlation between end-user interpretability of intermediate traces and final solution

Table 1: An example from the CoTemp QA dataset showing the outputs of Qwen3-1.7B and Llama-3.2-1B-Instruct models under different query setting. Correct final solutions are shown in green, and incorrect final solutions are shown in red. Correct intermediate traces are shown in blue, and incorrect intermediate traces are shown in red.

Model	Query Setting	Example
		Input Prompt: 'Answer the question based on the context: [\'Morus Hasratyan worked for The Newcastle upon Tyne Hospitals NHS Foundation Trust from September 11, 1972 to December 18, 1974.\', \'Morus Hasratyan is a member of the Communist Party of the Soviet Union in 1955.\', \'Morus Hasratyan works for Haigazian University from 1965 to 1966.\', " "Morus Hasratyan worked for Bishop\'s University from 1972 to 1975."", \'Morus Hasratyan worked for ISCTE – Lisbon University Institute from June, 1957 to December, 1960.\', \'Morus Hasratyan works for History Museum of Armenia from 1964 to 1975.\', \'Morus Hasratyan worked for Royal Air Force College Cranwell in February, 1959.\', \'Morus Hasratyan worked for University of Detroit Mercy in September, 1963.\', \'Morus Hasratyan worked for Tagesspiegel from May, 1957 to November, 1957.\', \'Morus Hasratyan worked for North Carolina State University in May, 1962.\'] Question: While Morus Hasratyan was working for Haigazian University, which employer did Morus Hasratyan work for during the same time period? Only return the answer.
		Gold Answer: "['History Museum of Armenia']" ; **Problem Category:** 'during' ; **Relevant Text:** "['Morus Hasratyan works for History Museum of Armenia from 1964 to 1975.']"
Qwen3-1.7b	Prompt	"Bishop's University"
	SFT - Vanilla	"['History Museum of Armenia']"
	SFT w/ Correct Traces	"<think>The temporal relation between the event in question and the event in context is: equal. I need to use the following facts to answer the question: ['Morus Hasratyan works for History Museum of Armenia from 1964 to 1975.']</think> <answer>['History Museum of Armenia']</answer>"
	SFT w/ Incorrect Traces	"<think>The temporal relation between the event in question and the event in context is: equal. I need to use the following facts to answer the question: Morus Hasratyan worked for ISCTE - Lisbon University Institute from June, 1957 to December, 1960.</think> <answer>['History Museum of Armenia']</answer>"
Llama-3.2-1B-It	Prompt	"Bishop's University"
	SFT - Vanilla	"['History Museum of Armenia']"
	SFT w/ Correct Traces	"<think>The temporal relation between the event in question and the event in context is: during. I need to use the following facts to answer the question: ['Morus Hasratyan works for History Museum of Armenia from 1964 to 1975.']</think> <answer>['History Museum of Armenia']</answer>"
	SFT w/ Incorrect Traces	"<think>The temporal relation between the event in question and the event in context is: overlap. I need to use the following facts to answer the question: Morus Hasratyan worked for Royal Air Force College Cranwell in February, 1959.</think> <answer>['History Museum of Armenia']</answer>"

accuracy, we also evaluate larger models (Qwen3-8B and Llama-3.1-8B) to study the effects of SFT with more complex traces. Due to practical constraints of human-subject studies, these experiments are conducted on the CoTemp QA domain (§4.1).

4.3.1 Reasoning Trace Generation

We 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 (prompts shown in §A.2).

4.3.2 Human-Subject Study

We conducted four separate user studies to evaluate the end-user interpretability of the four types of reasoning traces. In each study, a set of 25 participants were hired on Prolific and 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 traces. An example of the four types of traces can be found in §??. We use a between-subjects design to avoid bias from having participants compare multiple trace types themselves. We specifically test the following hypotheses:

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.

Table 2: CoTemp QA Results

Model	Query Setting	Final Solution Evaluations				Intermediate Trace Evaluations		
		Accuracy	F1	Precision	Recall	Classification Step Accuracy	IR Step Accuracy	Avg Trace Length (# tokens)
Qwen3-1.7b	Prompt	6.35	11.35	14.33	10.1	-	-	-
	SFT - Vanilla	60.33	74.88	82.15	71.3	-	-	-
	SFT - Correct Trace	52.88	70.63	79.45	66.33	**47.06**	**78.99**	45.8
	SFT - Incorrect Trace	**63.88**	**76.5**	**82.58**	**73.5**	20.36	56.92	34.15
Llama-3.2-1B-It	Prompt	7.48	13.78	17.58	12.15	-	-	-
	SFT - Vanilla	44.65	61.08	69.53	56.58	-	-	-
	SFT - Correct Trace	39.55	56.83	65.83	52.5	**39.09**	**79.4**	43.51
	SFT - Incorrect Trace	**45.58**	**61.15**	**69.65**	**57.23**	18.8	73.62	40.28

Table 3: Microsoft MARCO QA and Facebook bAbI QA Results

Model	Query Setting	Microsoft MARCO QA				Facebook bAbI QA			
		Avg Final Solution Accuracy (%)	Avg Trace Acc (Classification Step) (%)	Avg Trace Acc (IR Step) (%)	Avg Trace Length (# tokens)	Avg Final Solution Accuracy (%)	Avg Trace Acc (Classification Step) (%)	Avg Trace Acc (IR Step) (%)	Avg Trace Length (# tokens)
Qwen3-1.7B	Prompt	0	-	-	-	0	-	-	-
	SFT - Vanilla	3.4	-	-	-	**97.9**	-	-	-
	SFT - Correct Trace	**26.3**	**60.4**	40.6	68.14	94.41	**60.64**	**24.73**	43.25
	SFT - Incorrect Trace	20.3	6.9	**52.5**	85.07	95.21	17.82	0	42.45
Llama-3.2-1B-It	Prompt	1.7	-	-	-	12.8	-	-	-
	SFT - Vanilla	33.4	-	-	-	**96.5**	-	-	-
	SFT - Correct Trace	**33.7**	**59.9**	21.4	55.82	94.41	**61.7**	**24.73**	42.17
	SFT - Incorrect Trace	28.9	20	**43.9**	80.48	86.17	3.46	0	38.5

Each participant evaluated five fixed Q/A examples containing the question, predicted answer, and reasoning trace. After each example, participants rated the trace on a 5-point Likert scale for predictability, comprehensibility, interpretability, and faithfulness to context (Doshi-Velez and Kim, 2017; Jalali et al., 2023). To assess cognitive workload, we employed the NASA–TLX (Hart, 2006), focusing on mental demand, effort, and frustration. Further details on the user study, participant demographics, and procedures are provided in §B.

4.4 Implementation Details

Models were fine-tuned using the Hugging Face library (Wolf et al., 2020) on a single 80GB NVIDIA Tesla A100 GPU for 3 epochs (effective batch size 16, max sequence length 1024). We employed PEFT QLoRA (Dettmers et al., 2023) (rank 16, alpha 32) with a learning rate of 2e-4 (8-bit AdamW, cosine scheduler, 0.1 warm-up). Prompt experiments utilized vLLM (Kwon et al., 2023).

5 Results

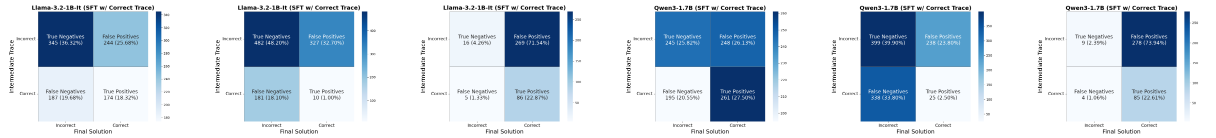
5.1 Final Solution & CoT Performance

As shown in Table 2, SFT with incorrect traces yields the highest final solution scores across Accuracy, Precision, F1, and Recall for both models. However, models SFT-ed with correct traces achieve higher Classification and IR Step accuracies due to the verifiability of their intermediate traces. Consistent patterns appear in the MARCO QA and bAbI QA datasets (Table 3): while final solution accuracy remains comparable, models trained with correct traces show stronger intermediate trace performance—particularly in bAbI QA, where those trained with incorrect traces perform poorly on the trace accuracy.

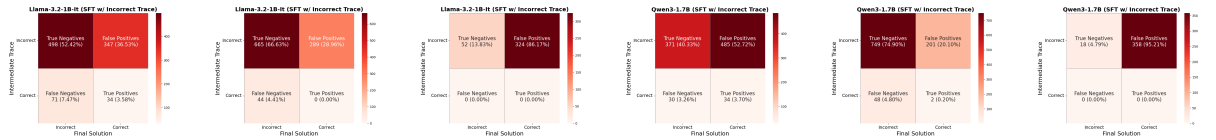
We hypothesize that the performance gain from incorrect traces arises from the model learning structural trace patterns while ignoring semantics. In our controlled setup, correct and incorrect traces share identical structures, and the SFT training with cross-entropy loss likely encourages the model to reproduce the paired <incorrect trace, correct final answer> during inference.

5.2 (Lack of) Correlation b/w Final Solution Accuracy & Trace Correctness

From the confusion matrices in Figure 2, the top row (Llama-3.2-1B-It) and bottom row (Qwen3-1.7B) both show a high rate of False Positives—cases where models produce correct answers but incorrect traces (25.7% in CoTemp, 32.7% in MARCO, and 71.54% in bAbI). In contrast, True Positives—both correct answers and traces—are few (e.g., 2.5% in MARCO and 22.61% in bAbI), indicating that many correct outputs rely on flawed intermediate reasoning. Figure 2 bottom row shows consistently high False Positive rates for



CoTemp (Llama, Correct) | MARCO (Llama, Correct) | bAbI (Llama, Correct) | CoTemp (Qwen, Correct) | MARCO (Qwen, Correct) | bAbI (Qwen, Correct)



CoTemp (Llama, Incorrect) | MARCO (Llama, Incorrect) | bAbI (Llama, Incorrect) | CoTemp (Qwen, Incorrect) | MARCO (Qwen, Incorrect) | bAbI (Qwen, Incorrect)

Figure 2: Confusion matrices showing Final Solution Accuracy (X-axis) vs Trace Accuracy (Y-axis) for the CoTemp QA, Microsoft MARCO QA, and Facebook bAbI QA datasets. **Top row:** SFT with Correct Traces on Llama-3.2-1B-It and Qwen3-1.7B models. **Bottom row:** SFT with Incorrect Traces on the same models.

both Llama-3.2-1B-It (top row) and Qwen3-1.7B (bottom row) across all datasets, indicating that finetuning on correct answers but incorrect traces enables strong final answer performance despite poor intermediate trace accuracy.

Tables 2 and 3 show varied outcomes across datasets: SFT with incorrect traces performs best on CoTemp QA, SFT with correct traces leads on MARCO QA, and SFT-Vanilla outperforms both on bAbI QA. These results collectively refute the assumption that semantically correct traces necessarily improve final solution performance.

To further strengthen our argument, we have also conducted a χ^2 statistical test with the null hypothesis: Trace accuracy and answer accuracy are independent. The test was done for both Qwen3-1.7B and Llama-3.2-1B-Instruct for the SFT w/ Incorrect trace setting. With a degree of freedom of 1, alpha=0.05, the critical value is 3.841 and we obtained χ^2 to be 0.34 and 2.93 (both < 3.841) for the two models respectively. Hence, the null hypothesis is correct.

5.3 Error Analysis b/w Final Solution & Intermediate Traces

Figure 2 shows that even when models were SFT-ed with correct traces and solutions, a substantial portion of cases featured correct traces preceding incorrect final answers. For Llama-3.2-1B-It, this occurred in up to 51.8% of CoTemp, 94.76% of MARCO, and 5.5% of bAbI samples; for Qwen3-1.7B, similar trends appeared (42.76%, 93.11%, and 4.49%, respectively). These findings indicate that training with correct intermediate traces does not reliably yield correct final predictions across datasets.

5.4 (Lack of) Correlation b/w Final Solution Accuracy & Trace Interpretability

5.4.1 SFT Evaluations

Figure 3 shows that, except for Qwen3-8B, SFT with R1 traces achieves the highest final accuracy, with the largest gain in Llama-3.2-1B-Instruct. In contrast, models trained with algorithmically generated semantically correct traces perform worst, even compared to those using R1 summaries or explanations. These results motivate a user study to assess the human interpretability of R1 traces.

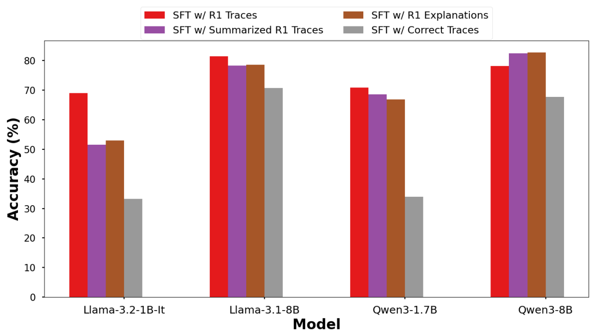


Figure 3: Final solution performance on CoTemp QA test dataset after SFT with different trace types on Llama and Qwen models.

5.4.2 Human-Subject Study Results

Table 4 shows that participants found algorithmically generated correct traces most interpretable across all dimensions, while R1 traces scored lowest. Summarized and explained R1 traces received

Table 4: Median participant ratings of reasoning traces across dimensions of end-user 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.

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

moderate ratings, suggesting improved comprehension by the subjects. R1 traces also imposed higher cognitive load, whereas correct traces caused less mental demand, effort, and frustration, making them easier to follow.

5.4.3 Statistical Analysis

We ran pairwise Mann–Whitney U tests with Bonferroni correction focusing on R1 vs. algorithmically generated correct traces for hypothesis testing with the following null hypotheses: **NH-1 (Interpretability):** There is no difference in interpretability ratings between R1 traces and algorithmically-generated correct reasoning traces. **NH-2 (Cognitive Workload):** There is no difference in cognitive workload ratings between R1 traces and algorithmically-generated correct reasoning traces.

Results show significant differences in both interpretability (predictability, comprehensibility, interpretability, faithfulness; all) and cognitive workload (mental demand, effort, frustration; all), leading us to reject both null hypotheses. Across all comparisons, R1 traces were consistently less interpretable and imposed higher cognitive load than other trace types.

5.5 Discussion

The key takeaways from our results can be summarized as follows:

1) *SFT w/ incorrect traces at times outperformed SFT w/ correct traces in final solution accuracy (§5.1).*

2) *Trace correctness did not guarantee final solution correctness. Solution correctness also did not imply a correct intermediate trace (§5.2 & §5.3).*

3) *Fine-tuning LLMs with the traces found to be the least interpretable by end-users led to the highest final solution accuracy, and vice-versa (§5.4).*

6 Conclusion

This work examines how an LLM's final solution accuracy relates to the semantic correctness and end-user interpretability of its intermediate traces after SFT. Using a knowledge distillation approach with verifiable problem decomposition on Open Book QA tasks, we find little correlation between trace and solution accuracy across Llama and Qwen models. Although SFT with R1 traces yields the best performance, user studies show these traces are the least interpretable, imposing higher cognitive load. This decoupling suggests that (1) verbose traces aid model reasoning more than human understanding, and (2) generating user-friendly explanations requires separate objectives or modules. Overall, we call for trace-based training methods that better balance LLM performance with human interpretability.

Limitations

OpenBook QA domain is not only representative of a critical application of today's interactive dialogue systems, but more importantly offers a controlled experiment environment for demonstrating the disconnect between trace fidelity and final accuracy. In other reasoning areas, as shown by commonly used test datasets like AIME/MATH500 for math and LiveCodeBench for code, there are many problems where creating a step-by-step reasoning process that works for every math or code problem is not possible. This creates a problem for evaluating the intermediate steps in these situations. The main point of this work is to demonstrate a gap between the accuracy of the steps and the final result for OpenBook QA problems, and the findings of this work are currently restricted to the aforementioned domains. These findings are also currently restricted by the choice and size of LLMs used in this work, and further efforts need to be carried to evaluate the extensibility of these results.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Perplexity AI. 2023. Perplexity ai. Accessed May 18, 2025.

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, and 1 others. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v2.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

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

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Google. 2023. Google gemini. Accessed May 18, 2025.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sandra G Hart. 2006. 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.

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

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

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. *arXiv preprint arXiv:2307.13702*.

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

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. 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).

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

Tyler McDonald and Ali Emami. 2024. Trace-of-thought prompting: investigating prompt-based knowledge distillation through question decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 397–410.

Microsoft. 2023. Microsoft copilot. Accessed May 18, 2025.

OpenAI. 2023. Chatgpt. Accessed May 18, 2025.

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

Nineta Polemi, Isabel Praça, Kitty Kioskli, and Adrien Bécue. 2024. Challenges and efforts in managing ai trustworthiness risks: a state of knowledge. *Frontiers in big Data*, 7:1381163.

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

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Scholkopf, and Mrinmaya Sachan. 2022. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Annual Meeting of the Association for Computational Linguistics*.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and 1 others. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.

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

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. 2025. 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.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

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, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shangzi Xue, Zhenya Huang, Jiayu Liu, Xin Lin, Yuting Ning, Binbin Jin, Xin Li, and Qi Liu. 2024. Decompose, analyze and rethink: Solving intricate problems with human-like reasoning cycle. *Advances in Neural Information Processing Systems*, 37:357–385.

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

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Additional Experiment Details

A.1 Dataset Distributions

Table 5: 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

Table 6: Train and Test data distribution for Microsoft MARCO QA dataset used in our SFT experiments.

Category	Train/Test Samples
description	1000 / 200
entity	1000 / 200
numeric	1000 / 200
location	1000 / 200
person	1000 / 200

Table 7: Train and Test data distribution for Facebook bAbI QA dataset used in our SFT experiments.

Category	Train/Test Samples
single-supporting-fact	200 / 20
two-supporting-facts	200 / 20
two-arg-relations	1000 / 100
counting	200 / 20
list-sets	200 / 20
conjunction	200 / 20
time-reasoning	200 / 20
basic-deduction	250 / 25
basic-induction	1000 / 100
positional-reasoning	125 / 12
size-reasoning	198 / 19

A.2 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}

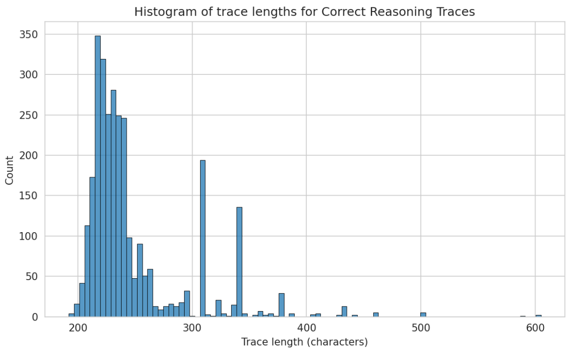
{R1 answer}

You have answered the question correctly. Please provide a detailed explanation of the reasoning behind your answer.

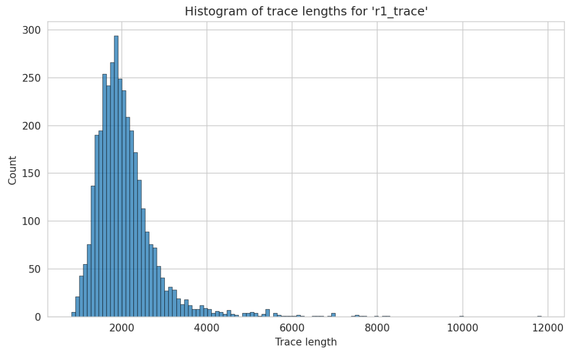
Explanation:

A.3 Trace Length Analysis

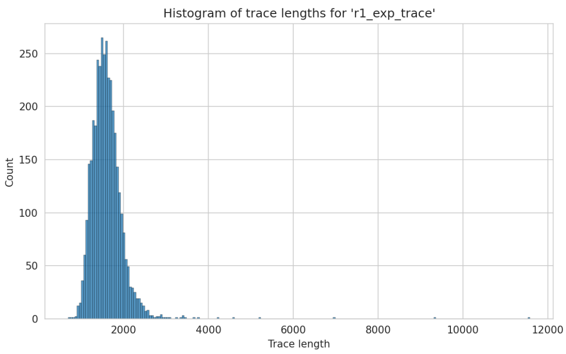
Figure 4 compares trace-length distributions for four reasoning trace types: algorithmically generated correct traces, R1 traces, post-hoc explanations of R1, and summarized R1 traces. It shows how R1 traces tend to be longer and more verbose, while summaries and explanations are more compact, affecting how much information users must process.



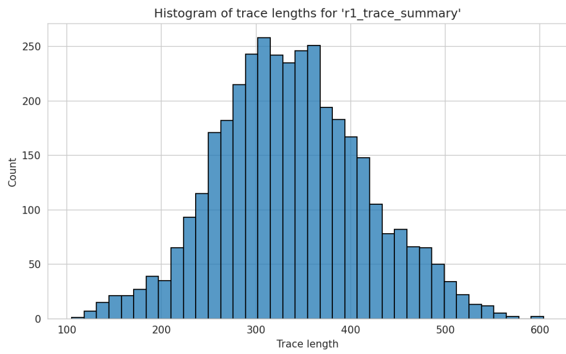
(a) Distribution of trace lengths of algorithmically generated correct traces.



(b) Distribution of trace lengths of R1 traces.



(c) Distribution of trace lengths of post-hoc explanations of R1 traces.



(d) Distribution of trace lengths of summaries of R1 traces.

Figure 4: Histogram comparison of trace lengths for the four reasoning trace types used in our study: (a) algorithmically generated correct traces, (b) R1 chain-of-thought traces, (c) post-hoc explanations of R1 traces, and (d) summarized R1 traces.

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 (Figure 5. The statement explained the goals of the study, what participants would be asked to do, and how their data would be handled.

User Study

What to Expect

This study has been reviewed and approved by the Institutional Review Board (IRB) with code:55564726. During the study, you will be presented with five passage-based questions. We expect the study to take approximately 15 minutes to complete.

Confidentiality

All data collected during this study will be kept strictly confidential. Your personal information will remain anonymous, and will only be accessible by authorized research investigators. The information will only be used for research purposes and will not be shared with any external entities.

Prolific ID

At the end of the study (when you click "Submit") you will be provided Prolific Remuneration Code. Please
retain that code for your reference and enter it in Prolific Interface to get the remuneration.

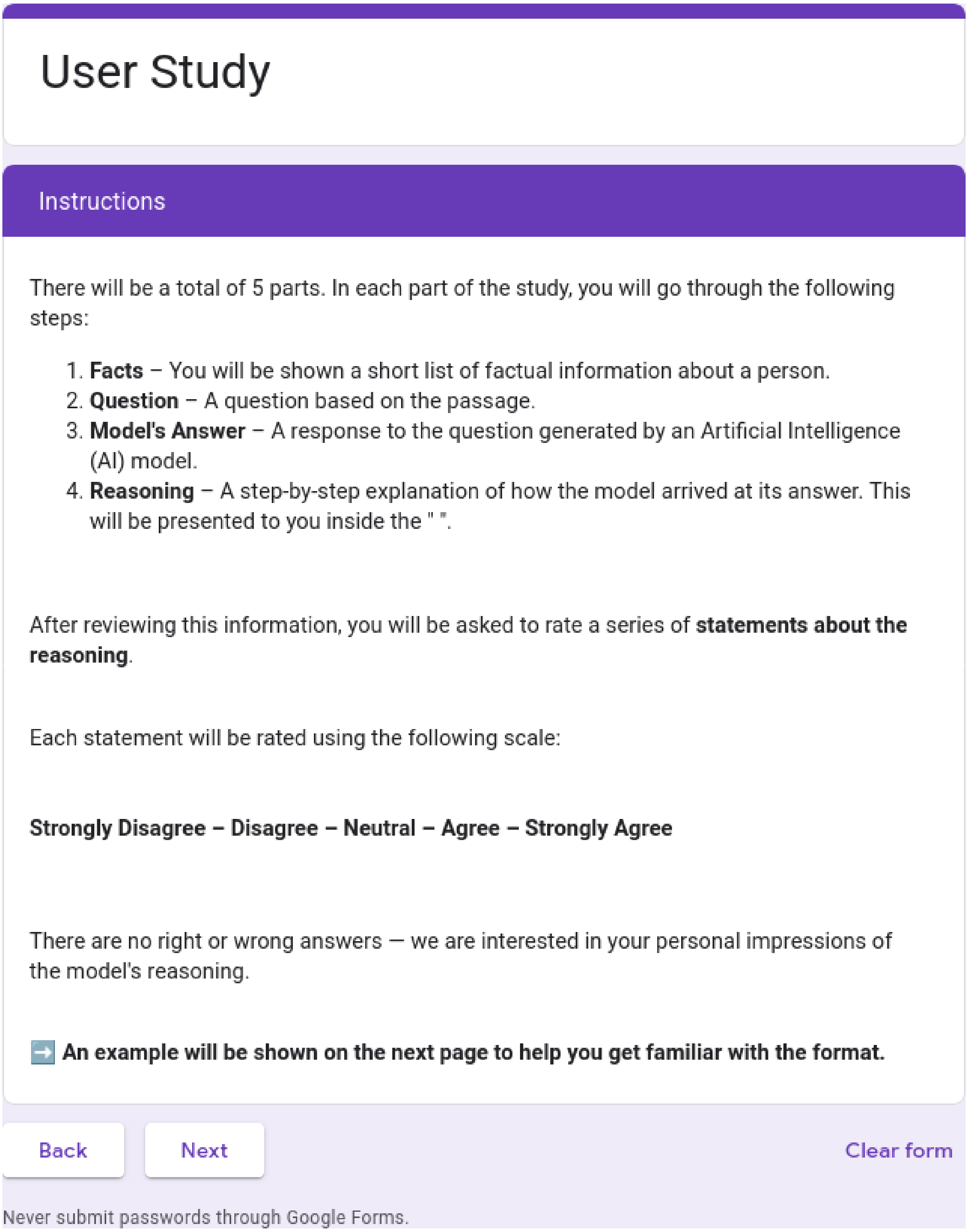
Figure 5: Consent statement shown to participants before starting the study.

B.3 Instructions

Participants were provided with detailed instructions describing the study structure (Figure 6). 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).



User Study

Instructions

There will be a total of 5 parts. In each part of the study, you will go through the following steps:

1. **Facts** – You will be shown a short list of factual information about a person.
2. **Question** – A question based on the passage.
3. **Model's Answer** – A response to the question generated by an Artificial Intelligence (AI) model.
4. **Reasoning** – A step-by-step explanation of how the model arrived at its answer. This will be presented to you inside the " ".

After reviewing this information, you will be asked to rate a series of **statements about the reasoning**.

Each statement will be rated using the following scale:

Strongly Disagree – Disagree – Neutral – Agree – Strongly Agree

There are no right or wrong answers — we are interested in your personal impressions of the model's reasoning.

➡️ **An example will be shown on the next page to help you get familiar with the format.**

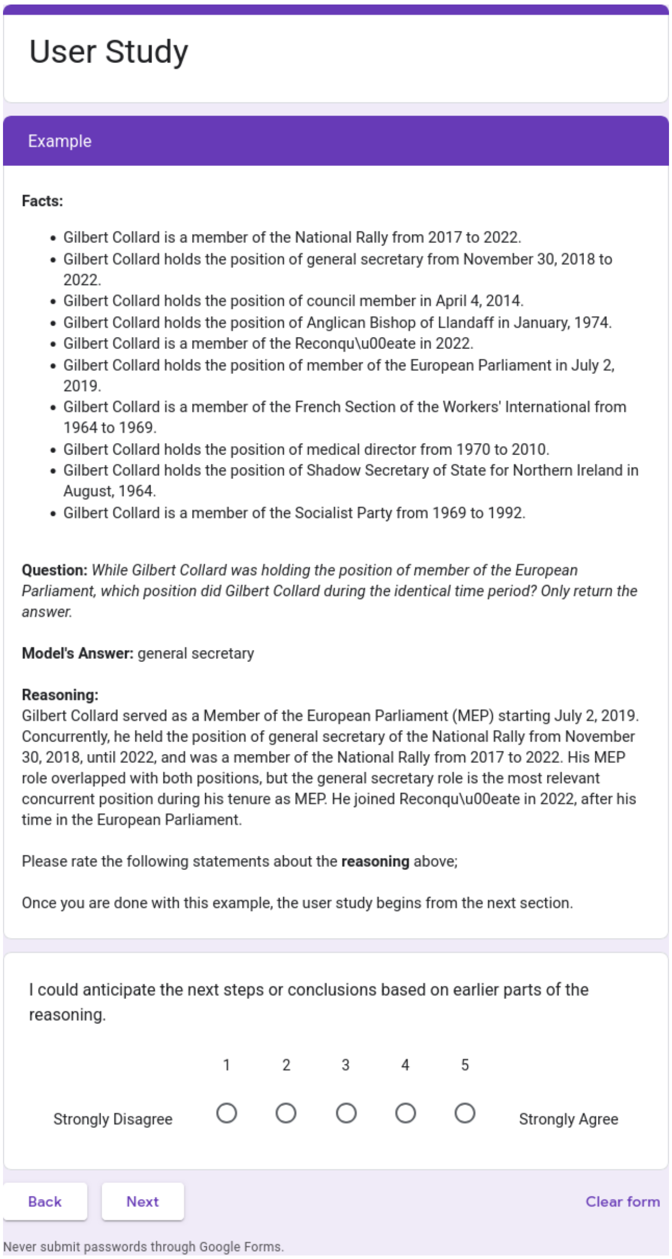
Back Next Clear form

Never submit passwords through Google Forms.

Figure 6: Instructions shown to participants before starting the study.

B.4 Q/A Task

Before beginning the main task, participants reviewed an example question and answer with reasoning (Figure 7). 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 (Figure 8).



User Study

Example

Facts:

- Gilbert Collard is a member of the National Rally from 2017 to 2022.
- Gilbert Collard holds the position of general secretary from November 30, 2018 to 2022.
- Gilbert Collard holds the position of council member in April 4, 2014.
- Gilbert Collard holds the position of Anglican Bishop of Llandaff in January, 1974.
- Gilbert Collard is a member of the Reconqu\u00eate in 2022.
- Gilbert Collard holds the position of member of the European Parliament in July 2, 2019.
- Gilbert Collard is a member of the French Section of the Workers' International from 1964 to 1969.
- Gilbert Collard holds the position of medical director from 1970 to 2010.
- Gilbert Collard holds the position of Shadow Secretary of State for Northern Ireland in August, 1964.
- Gilbert Collard is a member of the Socialist Party from 1969 to 1992.

Question: *While Gilbert Collard was holding the position of member of the European Parliament, which position did Gilbert Collard during the identical time period? Only return the answer.*

Model's Answer: general secretary

Reasoning:
Gilbert Collard served as a Member of the European Parliament (MEP) starting July 2, 2019. Concurrently, he held the position of general secretary of the National Rally from November 30, 2018, until 2022, and was a member of the National Rally from 2017 to 2022. His MEP role overlapped with both positions, but the general secretary role is the most relevant concurrent position during his tenure as MEP. He joined Reconqu\u00eate in 2022, after his time in the European Parliament.

Please rate the following statements about the **reasoning** above;

Once you are done with this example, the user study begins from the next section.

I could anticipate the next steps or conclusions based on earlier parts of the reasoning.

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Back Next Clear form

Never submit passwords through Google Forms.

Figure 7: Example shown to participants.

User Study

Question 1

Facts:

- Stine Bosse worked for Paramount Pictures from January 14, 1992 to November 24, 1999.
- Stine Bosse worked for IBM Almaden Research Center in October, 2008.
- Stine Bosse attended University of Copenhagen in 1987.
- Stine Bosse worked for Thomas Edison State University from June, 2007 to January, 2011.
- Stine Bosse works for Tryg from September 30, 2002 to February 1, 2011.
- Stine Bosse works for TDC from September 27, 2004 to February 28, 2006.
- Stine Bosse works for Alka from September 30, 2002 to February 1, 2011.
- Stine Bosse worked for National University of Science and Technology in January, 1999.
- Stine Bosse worked for Vassar College from March 26, 1998 to March 28, 1999.
- Stine Bosse works for Nykredit from January 16, 1989 to January 28, 1992.

Question: *While Stine Bosse was working for TDC, which employer did Stine Bosse work for within the same time interval? Only return the answer.*

Model's Answer: Tryg and Alka

Reasoning:
Stine Bosse worked for TDC from September 27, 2004, to February 28, 2006. During this period, she also worked for Tryg and Alka, both of which had overlapping employment dates from September 30, 2002, to February 1, 2011. Although the question asks for a singular employer, both Tryg and Alka qualify as valid answers.

Please rate the following statements about the **reasoning** above:-

Figure 8: Task shown to participants.