DO EXPLANATIONS GENERALIZE ACROSS LARGE REASONING MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large reasoning models (LRMs) produce a textual chain of thought (CoT) in the process of solving a problem. This CoT is potentially a powerful tool to understand the problem, surfacing a human-readable, natural-language explanation. However, it is unclear whether these explanations *generalize*, i.e. whether they capture general patterns about the underlying problem rather than patterns which are esoteric to the LRM. This is a crucial question in understanding or discovering new concepts, e.g. in AI for science. We study this generalization question by evaluating a specific notion of generalizability: whether explanations produced by one LRM induce the same behavior when given to other LRMs.

We find that CoT explanations do exhibit this form of generalization (i.e. they increase consistency between LRMs) and that this increased generalization is correlated with human preference rankings. We further analyze the conditions under which explanations do or do not yield consistent answers and propose a straightforward, sentence-level ensembling strategy that improves consistency. These results prescribe caution when using LRM explanations to yield new insights and outline a framework for characterizing LRM explanation generalization.

1 INTRODUCTION

The chains of thought (CoT) produced by large reasoning models (LRMs) have enabled strong performance on a range of complex tasks (Guo et al., 2025; Guha et al., 2025; Liu et al., 2025; Abdin et al., 2025; Agarwal et al., 2025). These CoT reasoning traces are often presented as human-readable explanations, but many researchers have questioned whether these traces can be made faithful to the true decision-making processes followed by LRMs (Barez et al., 2025; Chen et al., 2024; Shojaee et al., 2025; Xiong et al., 2025). In this paper, we examine a different issue that is related to faithfulness: we investigate the generalization of reasoning traces across different LRMs.

Our inquiry is motivated by the search for good natural-language explanations. For it is not good enough for an explanation to be correct; it is also important for an explanation to be *learnable*. That means: if we give the explanation to another person (or another agent), they should be able to understand it and draw the intended conclusions from it. In our setting, we quantify this question in terms of cross-LRM CoT generalization.

Importantly, models and humans may provide different explanations for the same concept, and uniformity is not our goal. Instead, we ask whether a given explanation, once produced, reliably guides other models to the same answer. This perspective reveals an underexplored dimension of CoT research: is a reasoning CoT produced by one LRM generalizable enough that it can lead a different LRM to follow the same reasoning, producing the same answer? Posing the question in this way enables an automated, quantitative evaluation of generalization for explanations, which has remained elusive despite generalization being a cornerstone of statistical machine learning. Practically, this question is critical for understanding whether an LRM explanation can provide usable insights into how a problem is solved. This question is especially critical in scientific discovery, where explanations that capture problem-level patterns, rather than model-specific quirks, could inspire novel human insights (Schut et al., 2025; Singh et al., 2024), especially as LRMs reach superhuman capabilities in domains such as science and mathematics (Wang et al., 2023; Romera-Paredes et al., 2024) and are increasingly used in educational settings (Kasneci et al., 2023; Bewersdorff et al., 2025).

We evaluate the effect of LRM explanations on improving the consistency between LRM answers in different ways (Fig. 1) across MedCalc-Bench. We find that LRM explanations do generalize, i.e. they increase consistency between LRMs (Fig. 1E), even improving consistency when the underlying explanation suggests a wrong answer. To further improve generalization, we propose a straightforward, sentence-level ensembling strategy that encourages the production of explanations less tied to the idiosyncrasies of any single model; we find that this further increases consistency between LRMs (Fig. 1E).

To evaluate the relationship between cross-model consistency and human preferences for explanations, we conduct a human study evaluating human preferences for various CoT explanations, suggesting that more consistent explanations may also be preferable for human users. Together, these results represent a step towards eliciting explanations from LRMs that are both transferable across models and informative to human users.

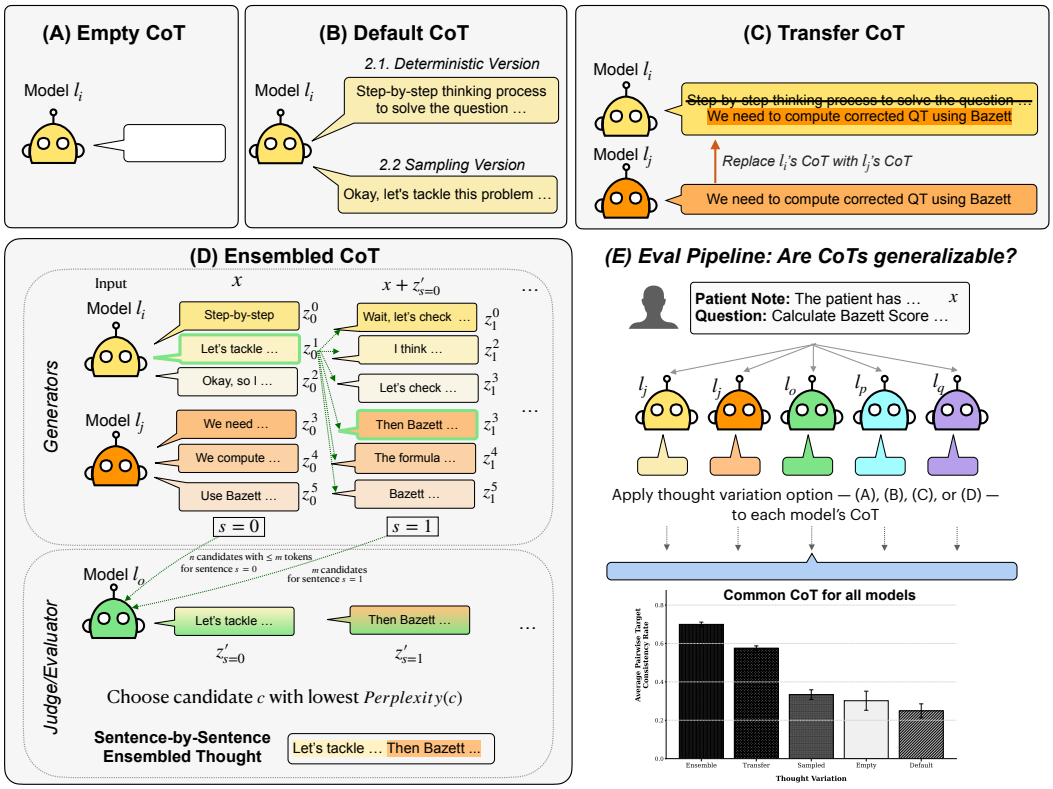

Figure 1: *Methods for eliciting reasoning chains and exploring generalization of Chain-of-Thought (CoT).* The figure illustrates four approaches to modifying or replacing model-generated CoTs. **Panel (E)** (bottom-right) shows how CoT generalization is evaluated for the MedCalc-Bench dataset across a set of models (A–E), where each model's reasoning can be substituted with one of the following variations: **(A) Empty CoT:** No reasoning text is provided between the model's thinking tags. **(B) Default CoT:** The model's own reasoning is used. (2.1) uses deterministic decoding, while (2.2) uses nucleus sampling (`do_sampling=True`). **(C) Transfer CoT:** Reasoning from one model is directly transferred to another, replacing its own. **(D) Ensembled CoT:** A generator–evaluator loop. Generator models produce $n = 3$ candidate sentences ($\leq$ 15 tokens each), forming k candidates. These are scored by the evaluator, and the least surprising candidate (lowest perplexity) is appended to the growing ensembled thought. This updated context is fed back into the generators, and the process repeats until an end-of-thought or maximum token limit is reached.

2 METHODS

Evaluating explanation generalizability Building on prior works that have focused on evaluating the faithfulness of an LRM explanation to a single model's reasoning, we evaluate the generalization of an explanation to a new LRM. Intuitively, a new LRM (trained using different techniques / datasets), should be a stand-in for a user simulator, allowing for evaluating whether an explanation is generalizable. Note, however, that this assumption may fail in the case that different LRMs share a common bias for a particular explanation.

Fig. 1 gives an overview of the different methods for eliciting and using LRM explanations that we consider. Concretely, given an LRM l_{gen} and a problem string x, we elicit a reasoning explanation by passing a model-specific prompt, that elicits a reasoning explanation $z = l_{\text{gen}}(x)$ before producing an answer $a = l_{\text{gen}}(x|z)$. For example, with the `Qwen/QwQ-32B` model (Team, 2025), we use a prompt of the form: {`Problem`}<think>{`Thinking Text`}</think>{`Answer`}....

When we test generalization, we supply the explanation from l_{gen}, i.e., $z = l_{\text{gen}}(x)$ to a different LRM l_{eval} amongst the population of models $L = \{l_i, l_j ... l_q\}$. We then produce an answer given an LRM l_{eval}, by giving it the explanation within the think tags. These explanations often contain answers. To ensure that the answer is not directly contained in the explanation, we further process the explanation by removing explicit answer declarations detected by an LLM.

We measure two metrics: accuracy A (measured via ground truth correctness) and consistency G (quantified by the frequency of matching responses). Concretely,

$$A = I(l_{\text{eval}}(x|l_{\text{gen}}(x)), a), \qquad G = \sum_{\substack{l \in L \\ z = l_{\text{gen}}(x)}} I(l_{\text{eval}}(x|z), l(x|z)) \tag{1}$$

where I is the scoring function specific to a dataset to see if two answers match.

Eliciting ensemble explanations Apart from extracting chain-of-thoughts from l_{gen} with various settings (empty, default, sampling), we also generate a chain-of-thought from a set of l_{gen}s, i.e, $L_{\text{gen}} = \{l_i, l_j, l_o ...\}$. Given this set of LRMs, we designate a subset as generators and a separate model as the evaluator. At each step, the generators produce n candidate sentences with m tokens ($n = 3$ and $m = 15$, in our case) conditioned on the context, which consists of the problem string x. The evaluator then selects the candidate with the lowest perplexity, which is appended to the ensembled chain of thought. The context is subsequently updated to include the original problem and all accumulated ensembled sentences. This sentence is part of z, which would be of size s, where size indicates the number of sentences we generate to create a complete ensembled thought. This process repeats until one of the generator models outputs an end-of-thought token or the maximum chain length ($m \cdot s$) is reached.

CoT variations As illustrated in Fig. 1, we evaluate four variations of CoT generation across models.

1. Empty CoT: The think text is an empty string, serving as a baseline method. Therefore, when the model generates its final answer, the preceding context is {`Problem`}<starting-think-tag>""<closing-think-tag>
2. Default CoT: The standard setting used in prior benchmarks, where the think text is generated by the model without modification. This method includes two sub-variations: one without sampling and one with sampling, covering both deterministic and non-deterministic default behaviors.
3. Transfer CoT: The think text is replaced by a default deterministic think text of another model. We test on various permutations of the models to see how different models' reasoning traces generalize across other models.
4. Ensembled Thoughts: The think text is replaced by explanations generated via Ensemble explanations.

For each thought type, we also present two versions — (1) with complete text, (2) with just hints and explanations, i.e., without answers. The intention behind this versioning is to understand to what

Table 1: LRMs used in this study. We focus on recent LRMs that have shown to be capable in various reasoning tasks.

Alias	Model	Huggingface ID	Citation
NRR	Nemotron-Research-Reasoning-Qwen-1.5B	`nvidia/Nemotron-Research-Reasoning-Qwen-1.5B`	(Liu et al., 2025)
OpenT	OpenThinker-7B	`open-thoughts/OpenThinker-7B`	(Guha et al., 2025)
OSS	gpt-oss-20b	`openai/gpt-oss-20b`	(OpenAI, 2025)
QwQ	Qwen/QwQ-32B	`Qwen/QwQ-32B`	(Team, 2025)
DAPO	DAPO-Qwen-32B	`BytedTsinghua-SIA/DAPO-Qwen-32B`	(Yu et al., 2025)

extent answers in chain of thoughts affect generalizability and accuracy. To remove answers, we use OpenAI's o4-mini model (OpenAI, 2025) as detailed in Appendix C.

3 Results

3.1 Experimental setup

Datasets To assess reasoning in a specialized and general domains, we adopt two benchmarks: We use `MedCalc-Bench` (Khandekar et al., 2024) to target medical domain-specific reasoning, and `Instruction Induction` (Honovich et al., 2022), which evaluates general reasoning capabilities. We extend the latter benchmark by incorporating 12 additional tasks to capture more complex general reasoning.

1. `MedCalc-Bench` (Khandekar et al., 2024): Each instance consists of a patient note and a question, which asks to compute a specific clinical value. We have a randomly chosen representation sample of size 100 of such calculation tasks.
2. `Instruction Induction` (Honovich et al., 2022): Each instance presents five input-output pairs and the model is tasked to generate a natural language instruction that captures their underlying relation. We evaluate this benchmark using 100 samples drawn from 20 tasks, with five examples per task.

In Eq. (1), the scoring function, I, for `MedCalc-Bench` is exact-matching while for `Instruction Induction` is BERTScore Zhang et al. (2020).

Models We deploy a series of LRMs; each listed in Table 1. We use huggingface implementations (Wolf et al., 2020) of each model in `bfloat16`.

User study We designed and conducted a user study with 15 participants to investigate whether greater generalizability correlates with users' perceptions of model CoT quality. The study was administered via Qualtrics, where participants received an anonymous survey link. Respondents, who were computer science and healthcare researchers, were instructed to evaluate reasoning chains from several model variations (kept anonymous to participants) across the following criteria: Clarity of Steps, Ease of Following, Confidence, followed by a Best Overall ranking. The questions were posed in Likert-scale manner:

- *Clarity of Steps:* The reasoning steps were clear and well explained (1 = Very unclear; 5 = Very clear)
- *Ease of Following*: The answer follows clearly from the reasoning steps. (1 = Very difficult; 5 = Very confident)

Table 2: Model configurations and reasoning approaches evaluated in the user study.

Reasoning Approach	Model Configuration
Deterministic Chain-of-Thought (CoT)	GPT-OSS-20B
Deterministic Chain-of-Thought (CoT)	DAPO-Qwen-32B
Ensemble CoT (Generator / Evaluator)	QwQ-32B + DAPO-Qwen-32B / GPT-OSS-20B
Ensemble CoT (Generator / Evaluator)	QwQ-32B + GPT-OSS-20B / DAPO-Qwen-32B

- *Confidence:* After reading, I feel confident I understood the reasoning. (1 = Not confident at all; 5 = Very confident)

The *Best Overall* ranking was asked in the following way: "Rank the following models' Chain-of-Thought explanations from most understandable to least understandable" (1 is the most understandable). While conducting this study, we did not collect demographic or any other personal identifying information. Because it would be unreasonable to expect participants to perform domain-specific tasks, our human evaluation focuses on whether the explanations are convincing or helpful, in contrast to the model-oriented accuracy and consistency metrics. All chain-of-thoughts were generated using `MedCalc-Bench` and the models and thought variations are summarized in Table 2. We constructed 10 examples, each paired with 4 chain-of-thoughts. For evaluation, participants were shown shown 5 examples, randomly selected and balanced across conditions.

3.2 Evaluating generalizability of LRM CoT explanations

Generalization without accuracy

LRM explanations generalize, even when the explanation induces an inaccurate answer.

Fig. 1 shows that the consistency rate among the models' conclusions increases in both transferred and ensemble settings. This includes 'incorrect' answers, where models arrive at the same wrong conclusion when provided with identical chains-of-thought. In Fig. 2, we measure how often a CoT leads models to converge on the same answer even when the predicted answer is not mentioned in the CoT (i.e., when the CoT mostly provides partial hints or reasoning). The left panel breaks down these "same answer" cases into convergence on the same correct versus the same incorrect answers. The right panel shows the proportion of consistently incorrect answers across CoT types. These results demonstrate that CoTs can systematically steer model reasoning, including toward the same conclusion, which indicates that CoT can exert a generalizing influence on model behavior even when the reasoning they provide can be incorrect. In Fig. 3, we report a breakdown of outcomes comparing CoT and baseline empty CoT reasoning across several scenarios, including cases where (a) an incorrect model prediction becomes correct after CoT transfer, (b) a correct prediction is perturbed, and (c) different forms of agreement or disagreement emerge across models. Regarding accuracy, Table 3 shows the effect of these chains-of-thought on different models. The takeaway is that models with weaker baseline accuracy can reduce the accuracy of other models when their CoTs are transferred even if the other model inherently performs better at their own baseline.

3.3 User Study

Human explanation preferences

LRM explanations that are more consistent receive higher user preference ratings.

Fig. 11 presents the statistical overview of our user study, showing box plots for each model and criterion. The plots display the mean (red diamond), median (red line), interquartile range (box boundaries), whiskers, and outliers (points). Overall, the default DAPO CoT and the ensembled QwQ+DAPO/OSS CoT were consistently perceived easier to understand than the default OSS CoT and the esembled QwQ+OSS/DAPO CoT.

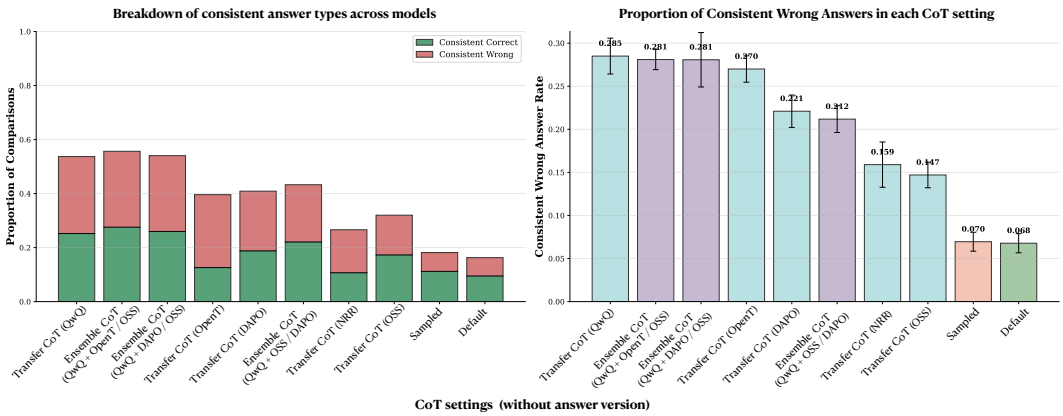

Figure 2: **Consistency breakdown across thought variations without answers.** *Left:* Proportion of consistent outputs separated into matching correct and matching incorrect conclusions. *Right:* Rate of consistent answers that are wrong across various thought settings.

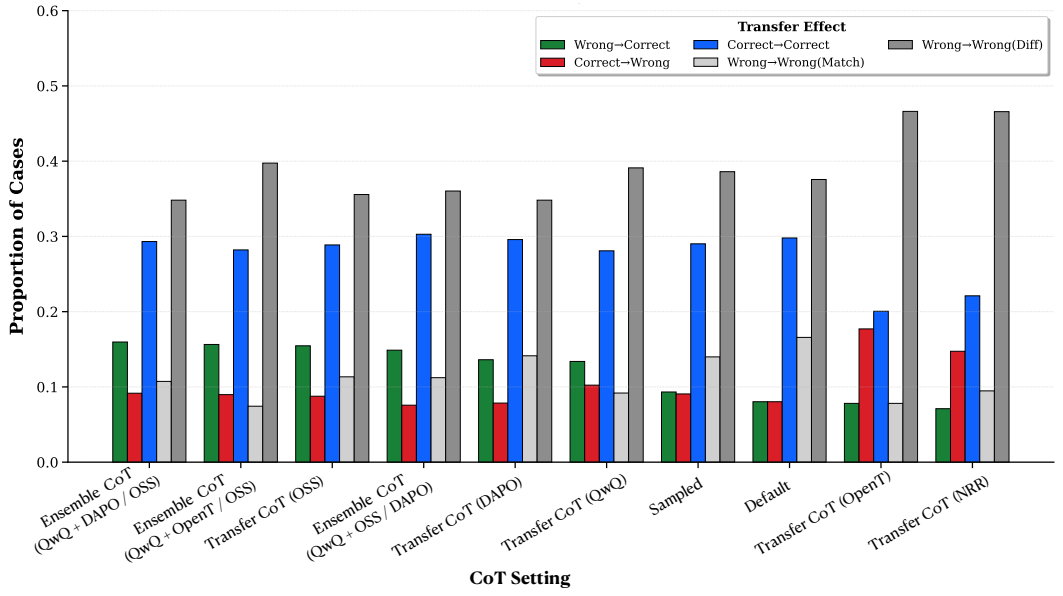

Figure 3: **CoT Transfer Effect Analysis** Distribution of transfer outcomes when Chain-of-Thought (CoT) reasoning is used or transferred across models. Each CoT setting is evaluated by comparing model predictions with CoT (which have explanations with answers removed) versus without CoT (empty baseline). Settings include, Default: models using their own generated CoT; Sampled: CoT generated through sampling; Transfer CoT l_{gen}: CoT transferred from model l_{gen} to all models; Ensemble CoT: combined CoT from multiple models. The five conditions represent: Wrong→Correct: cases where CoT successfully corrects errors (green); Correct→Wrong: cases where CoT misleads the model from correct to incorrect predictions (red); Correct→Correct: cases where CoT maintains correct predictions (blue); Wrong→Wrong(Match): both predictions incorrect with identical wrong answers (light gray); Wrong→Wrong(Diff): both predictions incorrect with different wrong answers (dark gray). Results are aggregated across multiple target models for each CoT setting. Settings are sorted by Wrong→Correct rate (descending).

Independent t-tests with Bonferroni correction confirmed that OSS was rated significantly worse than both DAPO ($p < 0.0001$) and QwQ+DAPO/OSS ($p < 0.0001$) in terms of *Clarity of Steps*. This pattern extended to *Ease of Following* and *Confidence*. In contrast, one ensemble variant per-

Table 3: Comparison of CoT accuracy across models for MedCalc-Bench and Instruction Induction

Method	Setting	MedCalc-Bench (Exact-Match)					Avg	Instruction Induction (BERTScore)					Avg
		NRR 1.5B	OpenT 7B	OSS 20B	QwQ 32B	DAPO 32B		NRR 1.5B	OpenT 7B	OSS 20B	QwQ 32B	DAPO 32B	
Empty CoT	No text	0.10	0.18	0.45	0.36	0.38	0.29	0.53	0.55	0.56	0.55	0.57	0.55
Default CoT	Full text	0.13	0.24	0.43	0.41	0.41	0.32	0.58	0.56	0.61	0.61	0.62	0.60
	W/o Ans	0.14	0.24	0.43	0.38	0.41	0.32	0.58	0.46	0.61	0.60	0.62	0.57
Sampled CoT	Full text	0.14	0.32	0.47	0.39	0.37	0.34	0.58	0.50	0.52	0.57	0.60	0.55
	W/o Ans	0.16	0.29	0.45	0.38	0.35	0.33	0.56	0.56	0.60	0.60	0.60	0.58
Trans. CoT (NRR)	Full text	0.13	0.13	0.21	0.22	0.29	0.20	0.58	0.56	0.60	0.58	0.62	0.59
	W/o Ans	0.14	0.15	0.24	0.30	0.34	0.23	0.58	0.57	0.60	0.60	0.63	0.60
Trans. CoT (OpenT)	Full text	0.24	0.26	0.24	0.26	0.27	0.25	0.60	0.57	0.43	0.62	0.62	0.57
	W/o Ans	0.21	0.24	0.26	0.26	0.25	0.24	0.60	0.46	0.59	0.59	0.61	0.57
Trans. CoT (OSS)	Full text	0.39	0.40	0.43	0.42	0.39	0.41	0.60	0.57	0.61	0.61	0.61	0.60
	W/o Ans	0.26	0.44	0.43	0.40	0.44	0.39	0.60	0.57	0.61	0.60	0.62	0.60
Trans. CoT (QwQ)	Full text	0.41	0.40	0.40	0.41	0.41	0.41	0.61	0.57	0.52	0.61	0.62	0.59
	W/o Ans	0.34	0.37	0.39	0.38	0.37	0.37	0.60	0.57	0.61	0.60	0.62	0.60
Trans. CoT (DAPO)	Full text	0.38	0.40	0.45	0.40	0.41	0.41	0.62	0.58	0.53	0.62	0.62	0.60
	W/o Ans	0.31	0.40	0.39	0.40	0.41	0.38	0.60	0.58	0.62	0.61	0.62	0.61
Ens. (QwQ+DAPO/OSS)	Full text	0.40	0.39	0.39	0.39	0.40	0.39	0.60	0.57	0.60	0.61	0.61	0.60
	W/o Ans	0.39	0.37	0.41	0.41	0.40	0.40	0.60	0.57	0.61	0.60	0.62	0.60
Ens. (QwQ+OSS/DAPO)	Full text	0.40	0.39	0.46	0.43	0.43	0.39	0.62	0.57	0.62	0.62	0.62	0.61
	W/o Ans	0.28	0.37	0.45	0.42	0.43	0.38	0.60	0.57	0.61	0.60	0.62	0.60
Ens. (QwQ+OpenT/OSS)	Full text	0.37	0.37	0.41	0.38	0.39	0.38	0.61	0.57	0.61	0.61	0.62	0.60
	W/o Ans	0.34	0.41	0.42	0.38	0.40	0.39	0.61	0.58	0.61	0.60	0.62	0.60

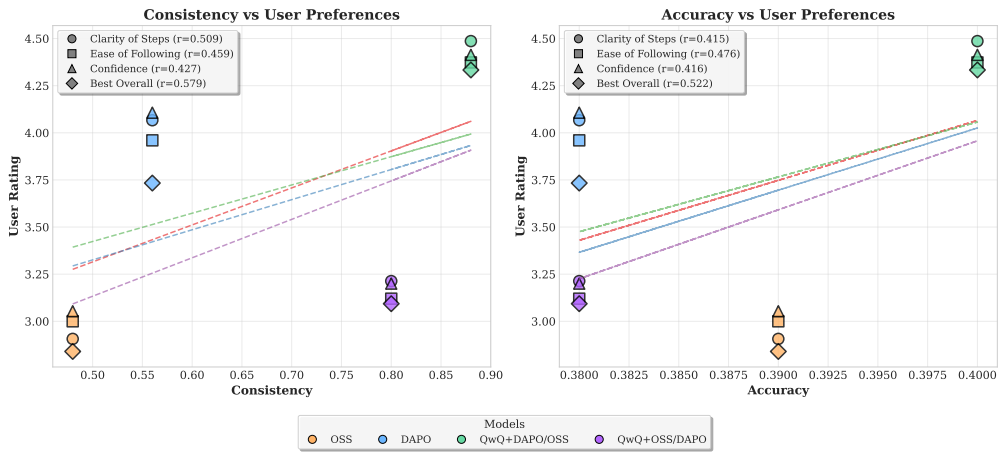

Figure 4: Scatter-and-trendline plots examining how model consistency (left) and model accuracy (right) relate to human user preferences. Each plot includes four user-rated dimensions: Clarity of Steps, Ease of Following, Confidence, and Best Overall. It also includes linear regression lines showing the strength and direction of their correlations. Consistency has a stronger and more visibly separable relationship with user preference than accuracy.

formed similarly to OSS, and the difference was not statistically significant ($p = 1.0$). All other comparison results, including those for *Best Overall*, showcased significant differences.

Between DAPO and ensembled QwQ+DAPO/OSS CoT, no significant differences emerged for *Clarity of Steps*, *Ease of Following* and *Confidence*. We also ran Wilcoxon signed-rank test and paired t-test. It yielded concordant results (see Table 4). The non-parametric and parametric tests agreed on significance for 23 out of 24 comparisons. For *Best Overall*, the ensemble was rated significantly higher ($p = 0.005$). These findings suggest that the ensemble with DAPO as generator and OSS as evaluator is the most effective configuration.

With Ensemble CoT + DAPO / OSS performing the best in Fig. 5 and in this study, the results support the claim that improving the generalizability of CoT can enhance perceived model quality explanation. In Fig. 4, we include a scatter plot illustrating this correlation. Notably, consistency appears to be a stronger predictor of user satisfaction than accuracy.

3.4 Analysis

Fig. 5 provides a detailed analysis of the average pairwise target consistency rate across different thought variations and configurations in both benchmarks. Pairwise target consistency is defined as the average consistency between pairs of models' outputs when provided with a chain of thought. We observe a clear increase in consistency from the default, empty, and sampled groups to the ensemble and transfer settings. Among these, the strongest performance comes from specific combinations of ensemble methods and thought variations in both MedCalc-Bench and Instruction Induction. In MedCalc-Bench, ensembles that use OSS as the evaluator achieve substantially higher consistency than other configurations. When OSS serves as the generator, consistency decreases, remaining above OSS on its own but closer to the lower end of the distribution. In Instruction-Induction, transferring OSS's CoT yields the strongest performance compared to other transfers. Similarly, the ensemble that uses OSS as the generator outperforms the other ensemble configurations. Taken together, these results suggest that models whose CoT transfers exhibit greater consistency also tend to function as more effective generators within ensemble transfers.

Fig. 6 examines how the average pairwise consistency rate changes depending on whether a pair includes the source model (i.e., the model whose CoT was used) or whether both models are targets that did not contribute to the CoT. This comparison highlights the extent to which similarity in responses persists when one member of the pair is the CoT creator versus when neither model generated the original chain of thought. Fig. 7 illustrate the models' self-consistency, which ranges from 20% to about 50% for respective models in question. Notably, these self-consistency levels differ from the cross-model consistency observed when models sample their responses.

4 Related work

Generating and improving natural-language explanations A large body of work extends chain-of-thought prompting (Wei et al., 2022) by probing or refining the explanations it produces. Examples include evaluating counterfactuals introduced into the chain of thought (Gat et al., 2023), testing their robustness to mistakes introduced into the reasoning chain (Lanham et al., 2023), or using contrastive CoT to induce reliance on the reasoning chain (Chia et al., 2023). A few works seek to improve the consistency in the generations made by an LLM, either between the generation and validation of LLMs (Li et al., 2023), between LLM predictions on implications of an original question (Akyürek et al., 2024), on counterfactual inputs for an original question (Chen et al., 2025b; Shihab et al., 2025), or by more generally introducing desirable structures into reasoning traces (Sun et al.). All of these methods can be used in conjunction with Ensemble explanations.

A similar line of work has studied generating explanations directly for a problem/dataset, rather than for a single example, e.g. describing distributions in natural language (Zhong et al., 2023; Singh et al., 2023) or human-readable programs (Romera-Paredes et al., 2024; Novikov et al., 2025). These works rely on some form of external verification for explanations (e.g. restricting an explanation to be python-runnable code) rather than allowing them to be flexible. A separate line of work has studied ensemble LLM generation (Tekin et al., 2024; Chen et al., 2025c), although not at the sentence-level and not for the purpose of explanation generation.

Assessing CoT explanations Model-generated text explanations have shown issues with faithfulness to the underlying LLM/LRM (Turpin et al., 2023; Ye & Durrett, 2022), e.g. LLM reasoning chains have been show to be inconsistent across counterfactuals (Mancoridis et al., 2025), sensitive to minor variations (Yeo et al., 2024), the answer may not follow from the chain (Xiong et al., 2025), may not reveal the info they really rely on (Chen et al., 2025a), inconsistently learn algorithms (Shojaee et al., 2025), can succeed at reasoning with invalid intermediate tokens (Stechly et al., 2025), or be trained to use dummy intermediate tokens (Pfau et al., 2024). Additionally, humans studies suggest that users perceive the wrong narratives from reasoning chains (Levy et al., 2025) and that users to not necessarily rank accurate reasoning traces for models higher (Bhambri et al., 2025a).

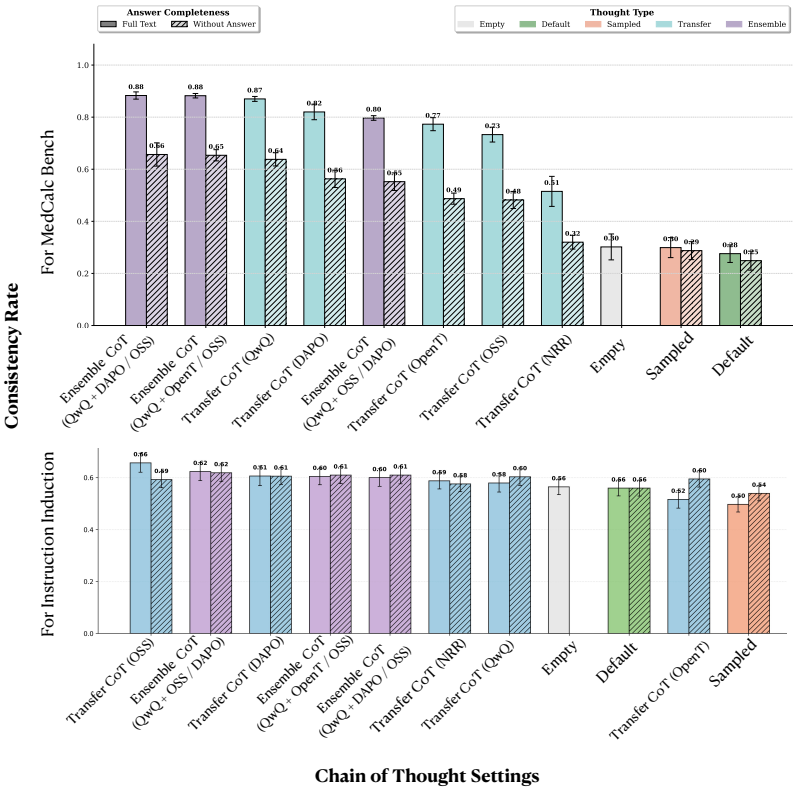

Figure 5: Average pairwise consistency across thought settings in MedCalc-Bench (above) and Instruction Induction (below). For thought variations indicating Ensemble CoT, models listed before the slash (/) serve as generators, while the model after the slash acts as the judge/evaluator. Results are reported both for the full text and for text with the final answer removed.

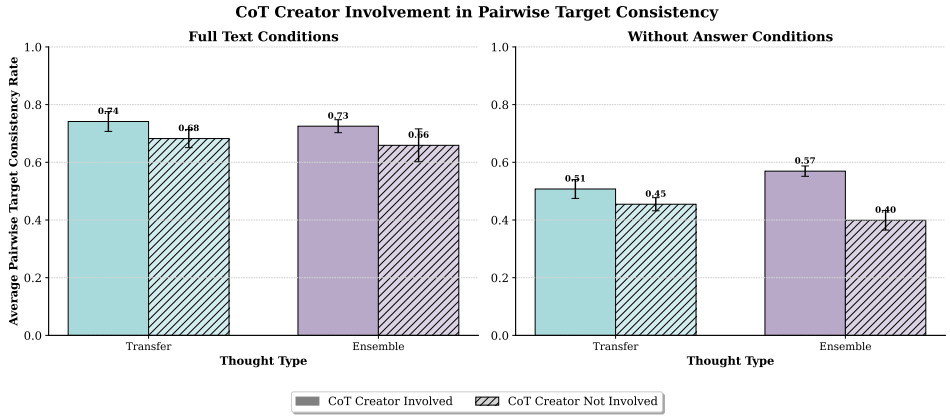

Figure 6: Average pairwise consistency for transfer and ensemble thoughts, comparing MedCalc Bench cases where a model is involved as a creator/source of the CoT versus cases where none of the models tested were part of the source.

See also other warnings about relying on LLM reasoning traces (Kambhampati et al., 2025; Bhambri et al., 2025b; Chua & Evans, 2025), including mechanistic analysis (Bogdan et al., 2025; Prakash et al., 2025), and on the difficulty of evaluating reasoning faithfulness (Zaman & Srivastava, 2025).

Evaluating natural-language explanations Prior works for evaluating natural-language explanations have aligned on one of three dimensions: consistency, plausibility, and faithfulness. *Consis-*

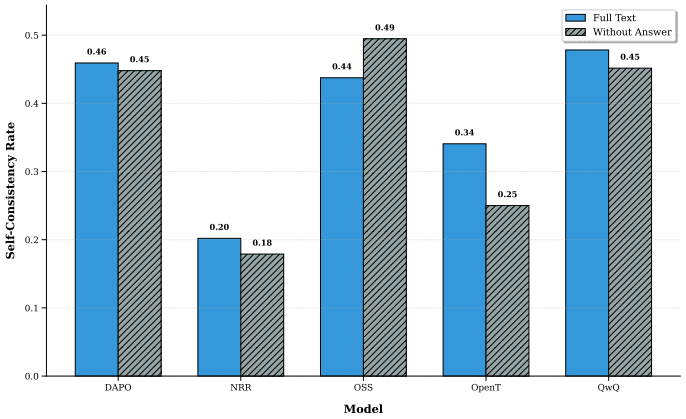

Figure 7: Model self-consistency rates for chain-of-thought generation on MedCalc-Bench. Bars compare default decoding versus sampling-based generation strategies.

tency, which we focus on in this work, measures if the model generates consistent explanations on similar examples (Hase & Bansal, 2020; Chen et al., 2024). *Plausibility* evaluates humans' preference of an explanation based on its factual correctness and logical coherence (Herman, 2017; Lage et al., 2019; Jacovi & Goldberg, 2020). It is different from *faithfulness*, which measures whether an explanation is consistent with the model's internal decision process (Harrington et al., 1985; Ribeiro et al., 2016; Gilpin et al., 2018; Jacovi & Goldberg, 2020). More broadly, explanation evaluation frameworks such as (Doshi-Velez & Kim, 2017), (Ribeiro et al., 2016), and surveys such as (Zhou et al., 2021; Hoffman et al., 2019) emphasize the distinction between human-centered and model-centered explanation quality and the need for metrics that reflect different goals of interpretability. We extend these metrics by evaluating explanations across models rather than within a single model. We measure whether a CoT from one LRM generalizes behaviorally to others through cross-model consistency. This provides a complementary perspective to existing explanation-evaluation frameworks focused on human preference or model faithfulness.

5 Discussion

In this work we have been motivated by the question, *"What is a good explanation?"* Unlike previous work that has focused on faithfulness or correctness, we have focused on the question of learnability: the notion that a good explanation should be effective at guiding a new student as the teacher intended. Taking advantage of the structure of LRMs as both producers and consumers of CoT, we have established a new framework for approaching this problem by measuring the generalization of CoT from one LRM to another.

Our work sets out a systematic way to conduct for an automated, quantitative evaluation of the generalizability of natural-language explanations, and we have demonstrated the use of our framework to measure a range of LRMs' ability to explain realistic tasks. We have shown that generalizability can be measured in terms of consistency of effects when transferring a CoT from one LRM to another. This consistency can be measured by removing explicit answers from the explanations, and measuring whether the other LRM arrives at the explained answer, right or wrong. We have also developed an ensembling method for generating highly-generalizable explanations.

Finally, our human study has validated that our consistency measure of generalization of chains-of-thought correlates with human preferences for better explanations. While our results here show some promise of LRM explanations at generalizing, the analyzed consistency metrics lay the groundwork for several interesting questions for future research, in particular raising the enticing possibility that LRMs could be trained to produce highly generalizable explanations. The framework of generalizable explanations also leads us to the question of whether the production of such explanations is itself a generalizable capability, and whether generalizable explanations produced by LRM can ultimately be used to help humans understand AI reasoning on subjects that extend human-produced training data, providing insights about knowledge beyond current human knowledge.

REPRODUCIBILITY

All experiments were run on workstations with 141GB NVIDIA H200 SXM GPUs using the HuggingFace Transformers library (Wolf et al., 2020). The codes and the dataset produced during this work will be made publicly available after publication.

ETHICS

Our work studies the generalizability of chain-of-thought (CoT) reasoning across different models and tasks. While CoT can improve performance and interpretability, its generalizability should be considered carefully. Reasoning patterns that transfer well in one setting may also reinforce shared mistakes in another, leading to consistent but incorrect outputs. In addition, reusing or combining CoTs across models may affect accuracy in ways that are not always predictable. These effects are particularly important to keep in mind in sensitive application areas, such as healthcare or law, where errors carry higher risks. We view this study as a step toward understanding both the benefits and limitations of CoT transfer. Future work should continue to explore when and how CoT generalizes reliably, and how to identify cases where it may not.

As for the user study we conducted, no personal information was collected during the user study experiments.

USE OF LARGE LANGUAGE MODELS

We used LLMs to help with plotting and minor editing of paper text.

REFERENCES

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

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925, 2025.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. Deductive closure training of language models for coherence, accuracy, and updatability. arXiv preprint arXiv:2401.08574, 2024.

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, pp. v2, 2025.

Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. Learning and Individual Differences, 118:102601, 2025.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Do cognitively interpretable reasoning traces improve llm performance? arXiv preprint arXiv:2508.16695, 2025a.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation, 2025b. URL https://arxiv.org/abs/2505.13792.

Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? arXiv preprint arXiv:2506.19143, 2025.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations. In Proceedings of the 41st International Conference on Machine Learning, pp. 7880–7904, 2024.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. arXiv preprint arXiv:2505.05410, 2025a.

Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. Towards consistent natural-language explanations via explanation-consistency finetuning. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 7558–7568, 2025b.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S Yu. Harnessing multiple large language models: A survey on llm ensemble. arXiv preprint arXiv:2502.18036, 2025c.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. Contrastive chain-of-thought prompting, 2023.

James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? arXiv preprint arXiv:2501.08156, 2025.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. ArXiv, 2017. URL https://arxiv.org/pdf/1702.08608.pdf.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box nlp models using llm-generated counterfactuals. arXiv preprint arXiv:2310.00603, 2023.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv preprint arXiv:1806.00069, 2018.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. arXiv preprint arXiv:2506.04178, 2025.

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.

Leo A Harrington, Michael D Morley, A Ščedrov, and Stephen G Simpson. Harvey Friedman's research on the foundations of mathematics. 1985. URL https://books.google.com/books/about/Harvey_Friedman_s_Research_on_the_Founda.html?id=2plPRR4LDxIC.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Proceedings of the Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.491.

Bernease Herman. The promise and peril of human evaluation for model interpretability. ArXiv, 2017. URL https://arxiv.org/pdf/1711.07414.pdf.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019. URL https://arxiv.org/abs/1812.04608.

Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. arXiv preprint arXiv:2205.10782, 2022.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.386.

Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! arXiv preprint arXiv:2504.09762, 2025.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences, 103:102274, 2023.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. Medcalc-bench: Evaluating large language models for medical calculations. Advances in Neural Information Processing Systems, 37:84730–84745, 2024.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. ArXiv, 2019. URL https://arxiv.org/pdf/1902.00006.pdf.

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.

Mosh Levy, Zohar Elyoseph, and Yoav Goldberg. Humans perceive wrong narratives from ai reasoning texts. arXiv preprint arXiv:2508.16599, 2025.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. arXiv preprint arXiv:2310.01846, 2023.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. arXiv preprint, 2025. URL https://arxiv.org/abs/2505.24864.

Marina Mancoridis, Bec Weeks, Keyon Vafa, and Sendhil Mullainathan. Potemkin understanding in large language models. arXiv preprint arXiv:2506.21521, 2025.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. Introducing openai o3 and o4-mini, 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: Sept 2025.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. arXiv preprint arXiv:2505.14685, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. URL https://doi.org/10.1145/2939672.2939778.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. Nature, 625(7995):468–475, 2024.

Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero. Proceedings of the National Academy of Sciences, 122(13):e2406675122, 2025.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Counterfactual sensitivity for faithful reasoning in language models. arXiv preprint arXiv:2509.01544, 2025.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2506.06941, 2025.

Chandan Singh, John X Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. Explaining data patterns in natural language with language models. In Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pp. 31–55, 2023.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. arXiv preprint arXiv:2402.01761, 2024.

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.

Chung-En Sun, Ge Yan, and Tsui-Wei Weng. Relif: A reliable, interpretable, and faithful lrm for trustworthy reasoning. In Mechanistic Interpretability Workshop at NeurIPS 2025.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. Llm-topla: Efficient llm ensemble by maximising diversity. arXiv preprint arXiv:2410.03953, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. ArXiv, 2023. URL https://arxiv.org/pdf/2305.04388.pdf.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. Nature, 620(7972):47–60, 2023.

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

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 (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. Measuring the faithfulness of thinking drafts in large reasoning models. arXiv preprint arXiv:2505.13774, 2025.

Xi Ye and Greg Durrett. Can explanations be useful for calibrating black box models? In Proceedings of the Association for Computational Linguistics, May 2022. URL https://aclanthology.org/2022.acl-long.429.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Kerem Zaman and Shashank Srivastava. A causal lens for evaluating faithfulness metrics. arXiv preprint arXiv:2502.18848, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. 2023.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5):593, 2021.

A SAMPLED DATA DISTRIBUTION

A.1 MEDCALC BENCH

We randomly sampled 100 data points from the `MedCalc-Bench` with seed 42 for our experiments. To show that these points are representative, we calculated the default deterministic CoT model performance across the sampled and full dataset. Figure 8 shows that the trend of best to worst model performance remains the same across default and across empty variations.

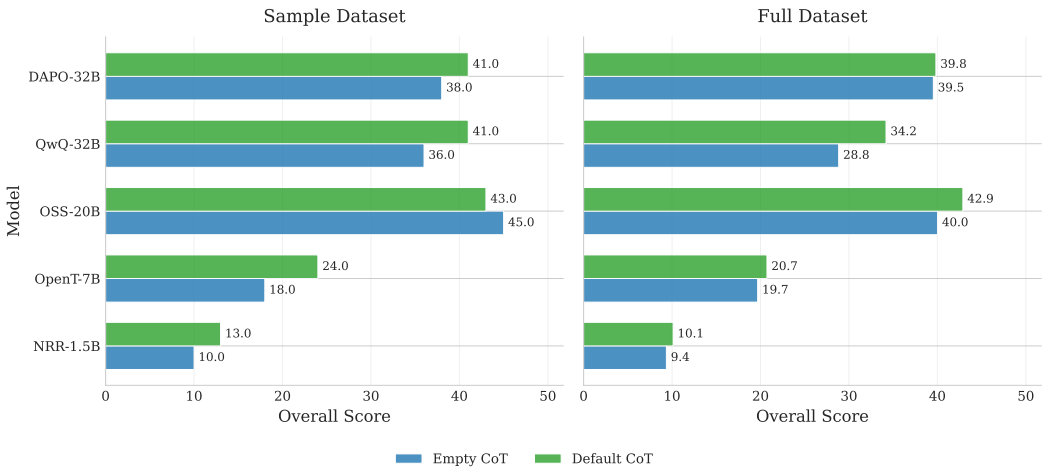

Figure 8: Model performance across all sampled and full data. The trends remain the same. Hence, the sample collected is a representative sample.

The model, `openai/gpt-oss-20b`, has three reasoning levels — low, medium, high. Based on the Medcalc benchmark, this model performs the best in the low level reasoning. Hence, for the rest of our experiments, we evaulated this model's CoT in low reasoning level effort.

B INSTRUCTION-INDUCTION

To create diverse and potentially complex instructions, we construct 12 new tasks in addition to the 24 tasks in the Instruction Induction Dataset (Honovich et al., 2022). The new tasks can be described as follows:

1. Reverse from middle: Locate the center point and reverse the left and right segments
2. Smallest Item Length: Find the shortest item and return its character count
3. Smallest even number square root: Identify the smallest even number and return its square root
4. Most vowel return consonant: Find the word with the most vowels and return its consonant count
5. Detect rhyme and rewrite: Detect rhyme schemes in poetry, then rewrite maintaining the same pattern.
6. Rank by Protein: Group foods into macronutrient categories and order by descending protein percentage
7. Translate to English: Recognize what language is being used and convert the main phrases into English
8. Square of Zodiac Animal: Find the zodiac animal in each list and output the square of its zodiac position
9. Alternate synonym antonym: Alternate between giving an antonym and synonym of the words in the sentence
10. Most consonant return vowel: Identify the word with the most consonants and return its vowel count

11. Identify fewest unique letters and return total letter count: Identify the word with the fewest unique letters and return its total letter count
12. First Word Alphabetically Return Reverse: Find the word that comes first alphabetically and return it in reverse

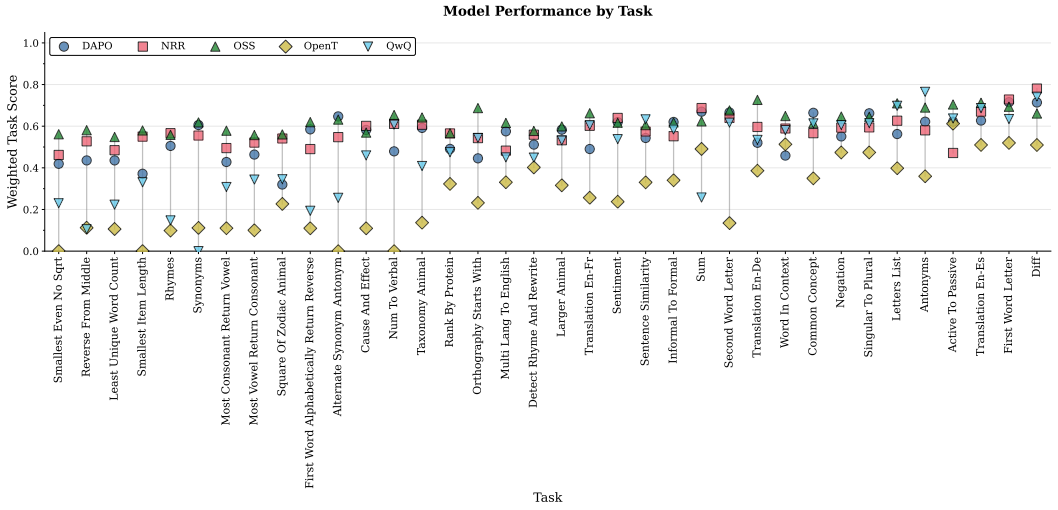

Figure 9: Model performance across all sampled data, including both the original instruction induction and newly introduced tasks.

C Remove Answers

We prompt OpenAI's o4-mini (OpenAI, 2025) with the content presented in Listing 1.

```
f"""Task: Keep only the hints from the text and remove answer sentences.

Definition:
- A "hint/explanation sentence" provides guidance that helps someone
  think about the problem without giving the final solution.
- An "answer sentence" directly states the final answer, solution,
  result, or conclusion.

Instructions:
1. Keep every hint/explanation sentence exactly as written.
2. Remove all answer sentences and statements.
3. Preserve the original wording, order, and formatting of the
   remaining text.
4. Do not add, rephrase, or generate any new text beyond what is
   already in the original.
5. Output only the hints.

Original text:
{chain-of-thought}"""
```

Listing 1: User Prompt Content for Answer Removal

D Ensemble Stats

The ensemble chain-of-thought method involves multiple models generating candidate sentences, after which judge or evaluator models select one candidate based on what they are least surprised by seeing, i.e., perplexity. The next sentence is then generated sequentially based on the context of

the question and the previous selected candidates, continuing this process iteratively. In Figure 10, we look at the distribution of candidates chosen from different pairs of generator models in various combinations of evaluated ensembles.

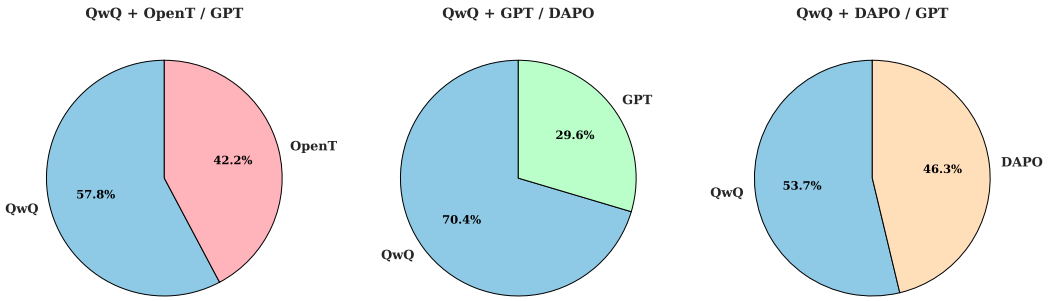

Figure 10: Proportion of selected candidate sentences from different generator models used to construct ensemble chain-of-thoughts across settings in MedCalc-Bench.

E More details on User Study

Fig. 11 shares a detailed overview of user study scores. We also share the detailed results of ilcoxon signed-rank test and paired t-test Table 4.

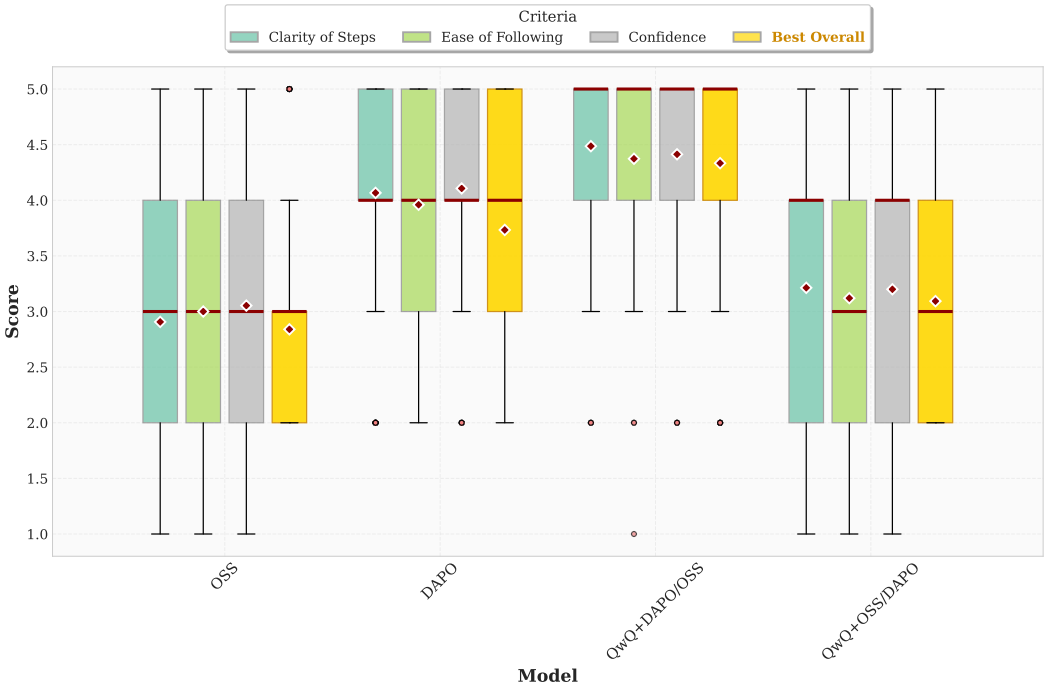

Figure 11: **User study results: Model performance comparison across evaluation criteria.** Box plots summarize the distribution of CoT evaluation scores for different models and model combinations. The three criteria are **Clarity of Steps** (leftmost box for each model), **Ease of Following** (second left box for each model), and **Confidence** (second right box for each model). Each box shows the median (red line), interquartile range (box boundaries), whiskers, and outliers (points), with red diamonds marking mean values. The **Best Overall** rankings (rightmost box for each model) report perceived understandability. Higher scores for all are better. The models are: OSS, DAPO, QwQ + DAPO/OSS and QwQ+OSS/DAPO.

Table 4: Pairwise Significance Matrix: Mean Differences (Wilcoxon Test)

	OSS	DAPO	QwQ+DAPO/OSS	QwQ+OSS/DAPO
Clarity of Steps				
OSS	–	$-1.16 * **$	$-1.58 * **$	$-0.31 * **$
DAPO		–	$-0.41 * **$	$+0.85 * **$
QwQ+DAPO/OSS			–	$+1.26 * **$
QwQ+OSS/DAPO				–
Ease of Following				
OSS	–	$-0.96 * **$	$-1.37 * **$	-0.12
DAPO		–	$-0.41 * **$	$+0.84 * **$
QwQ+DAPO/OSS			–	$+1.25 * **$
QwQ+OSS/DAPO				–
Confidence				
OSS	–	$-1.05 * **$	$-1.36 * **$	$-0.15*$
DAPO		–	$-0.31 * **$	$+0.91 * **$
QwQ+DAPO/OSS			–	$+1.21 * **$
QwQ+OSS/DAPO				–
Best Overall				
OSS	–	$-0.89 * **$	$-1.49 * **$	$-0.25 * **$
DAPO		–	$-0.60 * **$	$+0.64 * **$
QwQ+DAPO/OSS			–	$+1.24 * **$
QwQ+OSS/DAPO				–

Note: Values show mean differences (row model - column model). Green shading indicates row model rated lower (negative difference); red shading indicates row model rated higher (positive difference). Significance: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$. Gray indicates non-significant difference ($p > 0.05$). $n = 375$ pairs for all comparisons.