

DO EXPLANATIONS GENERALIZE ACROSS LARGE REASONING MODELS?

Koyena Pal^{1*}, David Bau¹, Chandan Singh²

¹Northeastern University ²Microsoft Research

{pal.k, d.bau}@northeastern.edu, chansingh@microsoft.com

ABSTRACT

Large reasoning models (LRMs) produce a textual chain of thought (CoT) in the process of solving a problem, which serves as a potentially powerful tool to understand the problem by 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 often exhibit this form of generalization (i.e. they increase consistency between LRMs) and that this increased generalization is correlated with human preference rankings and post-training with reinforcement learning. We further analyze the conditions under which explanations yield consistent answers and propose a straightforward, sentence-level ensembling strategy that improves consistency. Taken together, 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 (CoTs) 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 CoTs 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; Shoaee et al., 2025; Xiong et al., 2025). Another perspective of understanding the model’s chain-of-thought can be in terms of its utility as an explanation, i.e., not just whether it reflects the model’s internal reasoning, but whether it can effectively communicate that reasoning to other models and humans. Hence, in this paper, we examine a different property that is related to faithfulness: we investigate the generalization of reasoning traces across different LRMs.

Our study is motivated by the search for good natural-language explanations. For an explanation to be useful, it must not only be accurate, but also *understandable*; i.e. when presented to a new person (or LLM), the explanation should lead them to draw the intended conclusions from it. In our setting, we quantify this question in terms of cross-LRM CoT generalization. Specifically, we ask whether an explanation produced from one LRM reliably guides other LRMs to the same answer.

Posing the evaluation 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. This evaluation 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/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).

*Work done as part of the CBAI Fellowship

We evaluate the effect of LRM explanations on improving the consistency between LRM answers, when the explanation CoTs are generated in different ways (Fig. 1A-C). To further improve generalization, we propose a sentence-level ensembling strategy that encourages the production of explanations less tied to the idiosyncrasies of any single model and find that it increases consistency between LRMs (Fig. 1D). 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. Moreover, the Ensembled CoT improves consistency more than other strategies for eliciting a CoT.

We then evaluate the relationship between cross-model consistency and two other aspects. First, we conduct a human study evaluating human preferences for various CoT explanations, and find that more consistent explanations also appear preferable for human users. Second, we evaluate the relationship between cross-model consistency and post-training with reinforcement learning (RL), and find that RL post-training yields CoTs that are more consistent with other LRMs. Together, these results represent a step towards understanding when explanations from LRMs can be both transferable across models and informative to human users.

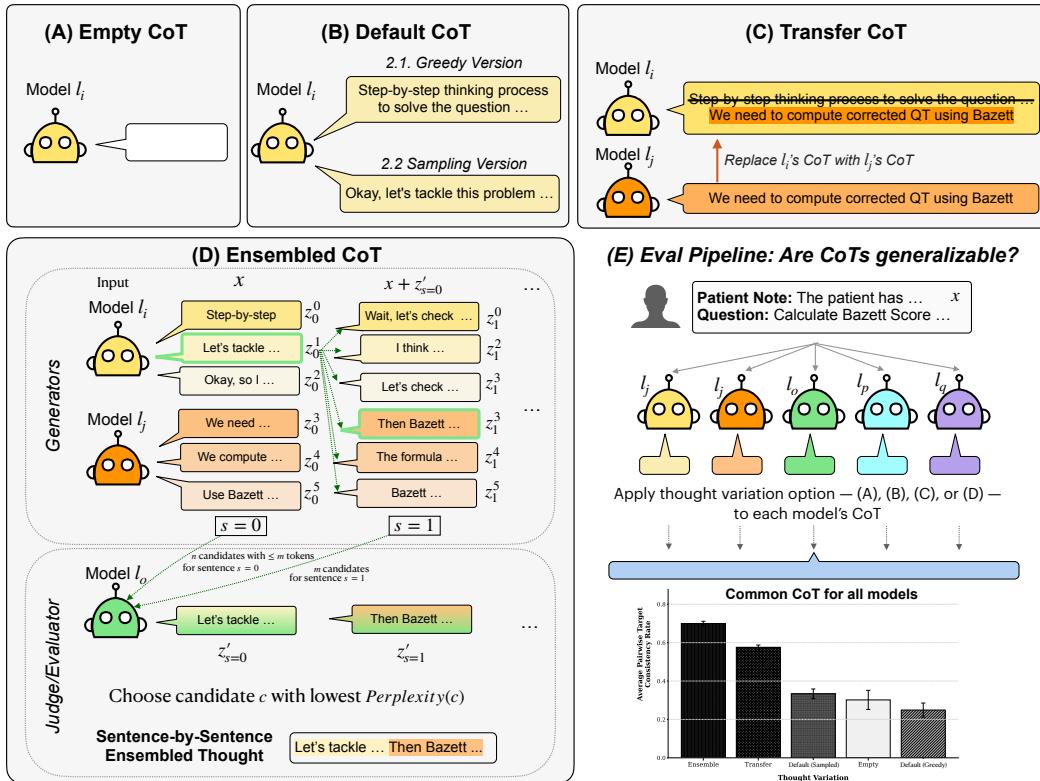


Figure 1: *Methods for eliciting chain-of-thought (CoT) explanations and evaluating them.* We evaluate explanation generation by querying LLMs for the answer to questions using CoTs in four different ways: **(A) Empty CoT.** No reasoning text is provided between the model’s thinking tags. **(B) Default CoT.** The model’s own reasoning is used. (2.1) uses greedy decoding, while (2.2) uses nucleus sampling. **(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 (≤ 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. **(E)** After eliciting CoTs in these four settings, we evaluate their generalization to new LRMs. The transfer CoT and Ensembled CoT significantly improve the consistency of the answer produced between the model producing the original CoT and the model producing the final answer (results shown here on the MedCalc-Bench dataset).

2 METHODS

Eliciting CoT explanations We seek to evaluate the generalization of a CoT explanation from one LRM to a new LRM. Fig. 1A-D gives an overview of the different methods for eliciting LRM explanations that we consider. Concretely, given an LRM l_{gen} and a problem string x , we elicit a CoT explanation by passing a model-specific prompt that elicits an 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} We evaluate four variations of CoT generation across models:

1. Empty CoT (Fig. 1A): 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 (Fig. 1B): 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 (Fig. 1C): 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 (Fig. 1D): The think text is replaced by explanations generated via Ensemble explanations. Given a 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 CoT. 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.

Evaluating CoT explanations Given the explanation from l_{gen} , i.e., $z = l_{\text{gen}}(x)$, we then test its generalization to a new LRM l_{eval} (see Fig. 1E). Specifically, we construct a prompt that includes the explanation within think tags, and then use it to query l_{eval} for an answer.

We measure two metrics: *Accuracy*, which measures the match between the generated answer and the ground truth answer a , and *Consistency*, which measures the match between answers from different models in a population of models $L = \{l_i, l_j \dots l_q\}$ when given the same explanation.

$$\text{Accuracy} = I(l_{\text{eval}}(x|l_{\text{gen}}(x)), a), \quad \text{Consistency} = \sum_{\substack{l \in L \\ z=l_{\text{gen}}(x)}} I(l_{\text{eval}}(x|z), l(x|z)) \quad (1)$$

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

One potential issue with this method of testing generalization is that explanations may directly include an answer, rather than an explanation leading to an answer. To avoid this issue, we process each explanation by prompting a separate LLM (gpt-04-mini, o4-mini-2025-04-16, OpenAI 2025) to remove any explicit answer declarations detected in the explanation (see details in Section A.2).

2.1 EXPERIMENTAL SETUP

Datasets and evaluation To assess reasoning in both specialized and general domains, we leverage two benchmarks. First, we use MedCalc-Bench (Khandekar et al., 2024), which targets medical domain-specific reasoning. Each instance consists of a patient note and a question, which asks to compute a specific clinical value, and we randomly select 100 examples of the dataset for evaluation. Second, we study Instruction Induction (Honovich et al., 2022), which evaluates general reasoning capabilities. Each instance presents five input-output pairs and the model

Table 1: LRMs used in this study. We focus on recent LRMs that have shown strong performance 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) |

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

| Reasoning Approach | Model Configuration |
|--------------------------------------|---------------------------------------|
| Default CoT (Greedy) | GPT-OSS-20B |
| Default CoT (Greedy) | 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 |

is tasked to generate a natural language instruction that captures their underlying relation. We randomly selected 8 of the original Instruction Induction tasks and extend the latter benchmark with 12 novel tasks to capture more complex general reasoning (see details in Section A.3). We evaluate the resulting 20 tasks by taking 100 random examples, five per task.

To evaluate results on these datasets, we specify the scoring function I from Eq. (1) to be exact-matching for MedCalc-Bench and BERTScore (Zhang et al., 2020) for Instruction Induction. We evaluate CoT generalization across the models in Table 1.

User study To investigate whether greater generalizability correlates with users’ perceptions of model CoT quality, we designed and conducted a user study. Participants were presented a CoT for a problem and asked to evaluate it across the following criteria¹ :

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

The study used 10 problems on the MedCalc-Bench dataset. For each problem, a CoT was generated in 4 different ways (see Table 2). For evaluation, participants were shown 5 examples, randomly selected and balanced across conditions. The study was administered via Qualtrics, where 15 participants, all of whom were computer science or healthcare researchers, received an anonymous survey link. Answers were given on a 5-point Likert scale. See more design details in Section A.7.

¹Note that the questions ask users to evaluate the explanations’ quality rather than directly use the explanation to perform a task; this limitation is largely because the MedCalc-Bench dataset requires significant domain expertise which users often did not have.

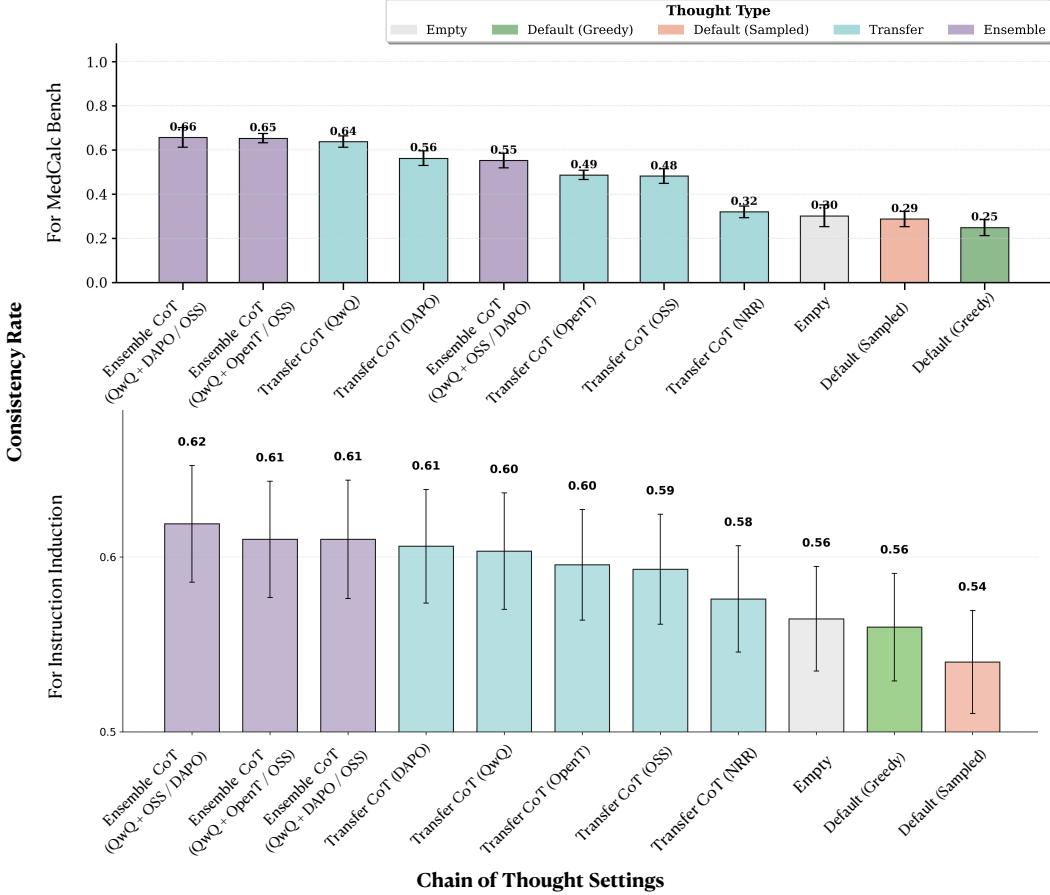


Figure 2: Average pairwise consistency across thought settings in MedCalc-Bench (above) and Instruction Induction (below). For thought variations indicating Ensemble CoT, models listed before the slash (/) serve as generators, while the model after the slash acts as the judge/evaluator.

3 RESULTS

3.1 EVALUATING GENERALIZABILITY OF LRM CoT EXPLANATIONS

Finding 1

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

Cross-model consistency Fig. 2 shows that providing LRMs with shared CoT explanations dramatically increases their consistency, i.e. how often they arrive at the same answer. Transfer and ensemble CoTs substantially improve consistency in both benchmarks. On MedCalc-Bench, consistency increases from a 25% baseline to 66%. Instruction-Induction shows similar gains, rising from 54% to 62% with ensemble CoT.

Interestingly, the default CoT with sampling outperforms greedy default CoT consistency in the MedCalc-Bench setting, which suggests that stochastic regularization may aid cross-model alignment for this calculation-heavy task. However, this advantage is not present in the Instruction-Induction setting, where sampled and greedy achieve comparable consistency rates. This suggests that the effectiveness of different decoding strategies for promoting cross-model consistency may depend on task characteristics.

Table 3: Comparison of CoT accuracy across models for MedCalc-Bench and Instruction Induction. See Section A.2 and Section A.5 for more details.

| Method | 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 | 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 (Greedy) CoT | 0.14 | 0.24 | 0.43 | 0.38 | 0.41 | 0.32 | 0.58 | 0.46 | 0.61 | 0.60 | 0.62 | 0.57 |
| Default (Sampled) CoT | 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) | 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) | 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) | 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) | 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) | 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) | 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) | 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) | 0.34 | 0.41 | 0.42 | 0.38 | 0.40 | 0.39 | 0.61 | 0.58 | 0.61 | 0.60 | 0.62 | 0.60 |

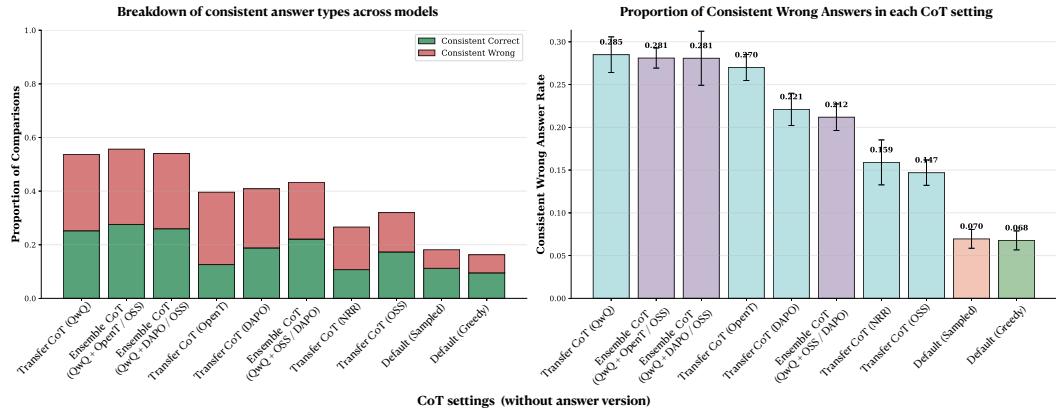


Figure 3: **Consistency breakdown across thought variations** *Left:* Proportion of consistent outputs separated into matching correct and matching incorrect conclusions. *Right:* Rate of consistent answers that are wrong across various thought settings.

The increase in consistency includes matching ‘incorrect’ answers, where models arrive at the same wrong conclusion when provided with identical chains-of-thought. Fig. 3 shows 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) in the MedCalc-Bench setting. 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.

Which LMs work best when extracting CoTs from an ensemble? 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 en-

semble 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.

Within-model prediction shifts across CoT conditions Fig. 4 shows 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.

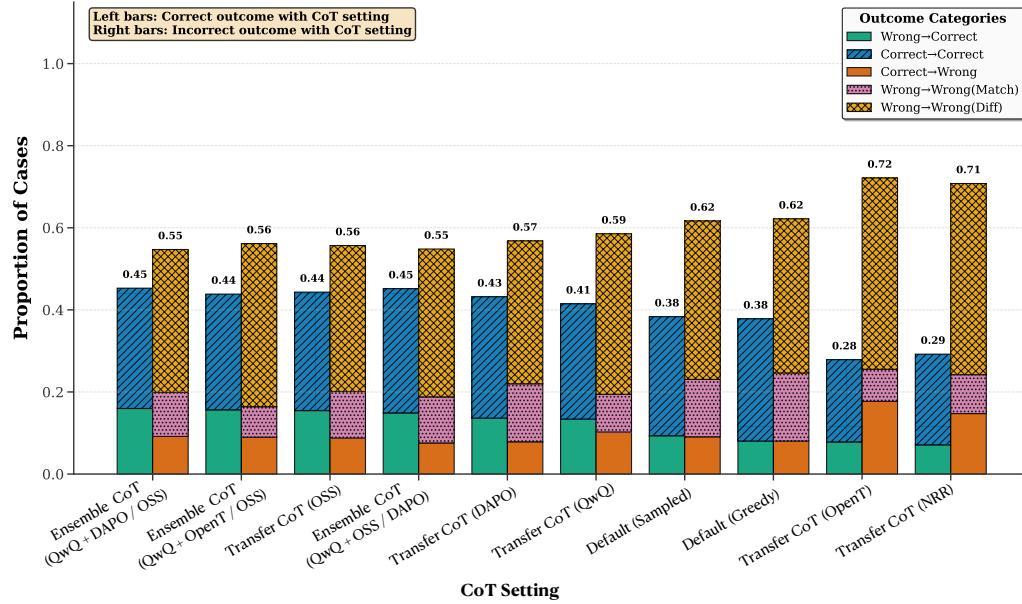


Figure 4: CoT Transfer Effect Analysis Distribution of transfer outcomes when CoT reasoning is used or transferred across models. Each CoT setting is evaluated by comparing model predictions with CoT versus without CoT (empty baseline). Settings include, Default: models using their own generated CoT; 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).

3.2 USER STUDY

Finding 2

LRM explanations that generalize to other LRMs receive higher user preference ratings.

The user study results suggest that improving the two metrics of CoT generalizability can enhance perceived model explanation quality. (see Fig. 5), although they should be interpreted with caution as they include only 4 LRM combinations. Of the two metrics, consistency appears to be a stronger predictor of user satisfaction than accuracy.

The user study results further show that the default DAPO CoT and the ensembled QwQ+DAPO/OSS CoT were consistently perceived as easier to understand than the default OSS

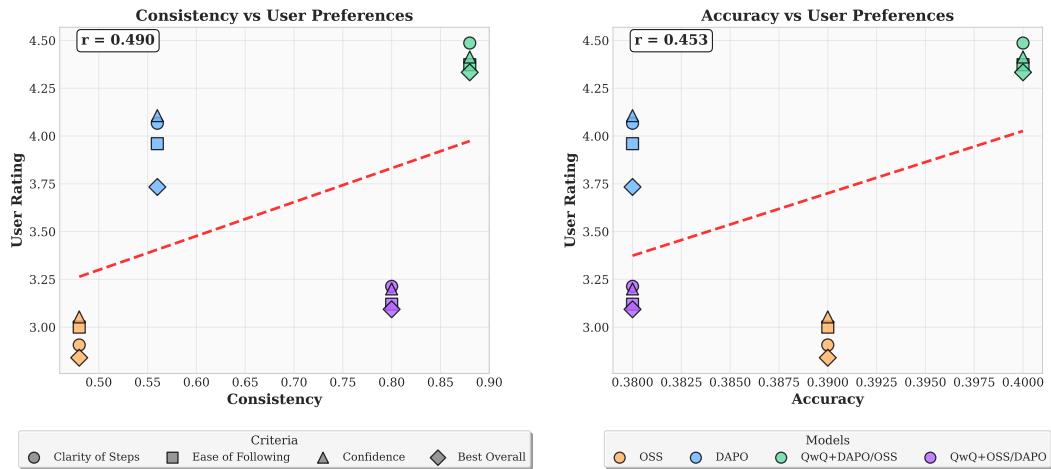


Figure 5: Comparing human user ratings of CoTs to LRM consistency (left) and LRM accuracy (right). Consistency and accuracy both show positive correlations with user ratings, with consistency showing a stronger trend. User ratings were collected for four criteria: Clarity of Steps, Ease of Following, Confidence, and Best Overall.

CoT and the ensembled QwQ+OSS/DAPO CoT. 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 performed 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*. For *Best Overall*, the ensemble was rated significantly higher ($p = 0.005$), suggesting that the ensemble with DAPO as generator and OSS as evaluator is the most effective configuration. See a further breakdown of results in Fig. A8.

3.3 IMPLICATIONS FOR RL TRAINING

Finding 3

RL post-training that improves an LRM’s performance improves its CoT consistency, but not necessarily its CoT transfer accuracy.

To assess the effect of reinforcement learning on model consistency and accuracy, we post-train two base models (Deepseek-R1-Distill-Qwen-1.5B and Llama-3.2-3B-Instruct) on the MedCalc-Bench dataset. We follow the RLVR-only setup and hyperparameters described for this dataset in (Lin et al., 2025).

Table 4: Effect of GRPO training on consistency and accuracy. Self-accuracy measures each model’s own performance. Avg. accuracy reports the mean accuracy when transferring each LRM’s CoT to five other models (NRR, OpenT, OSS, QwQ, DAPO). Avg. consistency measures mean pairwise consistency across the five models when evaluating each model’s transferred reasoning.

| Model | Self-Accuracy | Avg. Accuracy (Transfer CoT) | Avg. Consistency (Transfer CoT) |
|--------------------------------------|-----------------|------------------------------|---------------------------------|
| Deepseek-R1-Distill-Qwen-1.5B (Base) | 0.11 ± 0.03 | 0.30 ± 0.052 | 0.11 ± 0.007 |
| Deepseek-R1-Distill-Qwen-1.5B (GRPO) | 0.18 ± 0.04 | 0.29 ± 0.044 | 0.39 ± 0.061 |
| Llama-3.2-3B-Instruct (Base) | 0.12 ± 0.03 | 0.25 ± 0.050 | 0.25 ± 0.019 |
| Llama-3.2-3B-Instruct (GRPO) | 0.45 ± 0.05 | 0.44 ± 0.039 | 0.46 ± 0.035 |

Table 4 reveals a notable difference between consistency and accuracy improvements, after RL post-training. While both models show substantial consistency gains, their accuracy trajectories diverge: Deepseek’s average accuracy remains flat despite improved self-accuracy, while Llama shows coupled improvements across all metrics. This demonstrates that consistency and accuracy are separable properties that may vary independently. Future training methods may benefit from explicitly optimizing both consistency and transfer accuracy to enhance the generalizability of chain-of-thought reasoning across models.

4 RELATED WORK

Generating and improving natural-language explanations A large body of work extends CoT prompting (Wei et al., 2022) by probing or refining the explanations it produces. Examples include evaluating counterfactuals introduced into the CoT (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.).

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 shown 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, human studies suggest that users perceive the wrong narratives from reasoning chains (Levy et al., 2025) and that users do not necessarily rank accurate reasoning traces for models higher (Bhambri et al., 2025a). 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) and monitoring chain-of-thoughts (Korbak et al., 2025; Guan et al., 2025).

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

Our analysis is motivated by the question, “*What is a good explanation?*” Unlike previous work that has focused on faithfulness or correctness, we focus on the question of generalization: the notion that a good explanation should be effective at guiding a new user as the explainer intended. Taking advantage of the structure of LRMAs as both producers and users of CoT, we establish a framework for approaching this problem by measuring the generalization of CoT from one LRM to another.

While the focus here is on cross-LRM generalization, we note that this measure has some bearing on whether explanations generalize to humans (as seen in our human survey). As LRMs become increasingly adept at simulating a user (Dou et al., 2025; Naous et al., 2025), the same framework can help us evaluate an explanation’s generalizability to humans. However, currently this assumption may fail in the case that different LRMs share a common bias for a particular explanation.

The framework here lays the groundwork for a variety of future work. One important domain in AI safety may seek generalizable CoTs for the purpose of monitoring and understanding by LLM-based safety tools (Bedi et al., 2025; Zhao et al., 2025). Another line of work may seek to find CoTs that generalize across diverse examples for the purpose of generating new knowledge (Venkatraman et al., 2025; Qu et al., 2025; Feng et al., 2026). Finally, future work could explore improvements in explanation generation with ensembles (building on the straightforward strategy introduced here) or in improving RL post-training pipelines to incentivize generating generalizable explanations. We hope that properly generating and evaluating generalizable explanations produced by LRM can ultimately help humans understand subjects beyond current human knowledge.

ACKNOWLEDGMENTS

The authors thank Chris Ackerman, Lace Padilla, and Byron Wallace for advice on this project, as well as the members of the Baulab and Millicent Li for their feedback on the paper. We also thank the participants in our user study, whose contributions provided valuable insights for this work. This research was supported by the generous funding of the Cambridge-Boston Alignment Initiative (CBAI) Fellowship and Coefficient Giving.

REPRODUCIBILITY

All experiments were run either on workstations with 141GB NVIDIA H200 SXM GPUs or 80GB NVIDIA A100 GPUs using the HuggingFace Transformers library (Wolf et al., 2020). The code and dataset produced during this work is publicly available here: <https://genex.baulab.info>.

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. In 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 coding for plots 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.
- Afro 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.
- Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. Fidelity of medical reasoning in large language models. [JAMA Network Open](#), 8(8):e2526021–e2526021, 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>.
- Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants? In [Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing](#), pp. 35200–35278, 2025.
- Jean Feng, Avni Kothari, Patrick Vossler, Andrew Bishara, Lucas Zier, Newton Addo, Aaron Kornblith, Yan Shuo Tan, and Chandan Singh. Human-ai co-design for clinical prediction models, 2026. URL <https://arxiv.org/abs/2601.09072>.

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](https://arxiv.org/abs/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](https://arxiv.org/abs/1806.00069), 2018.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring monitorability, 2025. URL <https://arxiv.org/abs/2512.18311>.

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](https://arxiv.org/abs/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](https://arxiv.org/abs/2501.12948), 2025.

Leo A Harrington, Michael D Morley, A Šcedrov, 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](https://aclanthology.org/2020.acl-main.491), 2020. URL <https://aclanthology.org/2020.acl-main.491>.

Bernease Herman. The promise and peril of human evaluation for model interpretability. [ArXiv](https://arxiv.org/pdf/1711.07414.pdf), 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](https://arxiv.org/abs/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](https://aclanthology.org/2020.acl-main.386), 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](https://arxiv.org/abs/2504.09762), 2025.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. [Learning and individual differences](https://doi.org/10.1037/a00274), 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](https://proceedings.neurips.cc/paper/2024/084730-84745.pdf), 37:84730–84745, 2024.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL <https://arxiv.org/abs/2507.11473>.

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](https://arxiv.org/pdf/1902.00006.pdf), 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](https://arxiv.org/abs/2307.13702), 2023.
- Mosh Levy, Zohar Elyoseph, and Yoav Goldberg. Humans perceive wrong narratives from ai reasoning texts. [arXiv preprint arXiv:2508.16599](https://arxiv.org/abs/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](https://arxiv.org/abs/2310.01846), 2023.
- Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training llms for ehr-based reasoning tasks via reinforcement learning, 2025. URL <https://arxiv.org/abs/2505.24105>.
- Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prrol: Prolonged reinforcement learning expands reasoning boundaries in large language models. [arXiv preprint](https://arxiv.org/abs/2505.24864), 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](https://arxiv.org/abs/2506.21521), 2025.
- Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. [arXiv preprint arXiv:2510.06552](https://arxiv.org/abs/2510.06552), 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](https://arxiv.org/abs/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](https://openreview.net/forum?id=NikbrdtYVG), 2024. URL <https://openreview.net/forum?id=NikbrdtYVG>.
- Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Sharam, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. [arXiv preprint arXiv:2505.14685](https://arxiv.org/abs/2505.14685), 2025.
- Yuxiao Qu, Anikait Singh, Yoonho Lee, Amrith Setlur, Ruslan Salakhutdinov, Chelsea Finn, and Aviral Kumar. Rlad: Training llms to discover abstractions for solving reasoning problems. [arXiv preprint arXiv:2510.02263](https://arxiv.org/abs/2510.02263), 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](https://doi.org/10.1145/2939672.2939778), 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](https://doi.org/10.1038/nature25995), 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](https://doi.org/10.1073/pnas.2406675122), 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](https://arxiv.org/abs/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](https://arxiv.org/abs/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](https://doi.org/10.1145/3551438.3551457), 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](https://arxiv.org/abs/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](https://arxiv.org/abs/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 lilm ensemble by maximising diversity. [arXiv preprint arXiv:2410.03953](https://arxiv.org/abs/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](https://arxiv.org/pdf/2305.04388.pdf), 2023. URL <https://arxiv.org/pdf/2305.04388.pdf>.
- Siddarth Venkatraman, Vineet Jain, Sarthak Mittal, Vedant Shah, Johan Obando-Ceron, Yoshua Bengio, Brian R Bartoldson, Bhavya Kailkhura, Guillaume Lajoie, Glen Berseth, et al. Recursive self-aggregation unlocks deep thinking in large language models. [arXiv preprint arXiv:2509.26626](https://arxiv.org/abs/2509.26626), 2025.
- 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](https://doi.org/10.1038/nature1272), 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](https://proceedings.neurips.cc/paper/2022/0c9f65e043559f9a69d7484771d4453.pdf), 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](https://aclanthology.org/2020.emnlp-demos.6), 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](https://arxiv.org/abs/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](https://aclanthology.org/2022.acl-long.429), 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](https://arxiv.org/abs/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 lilm 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](https://arxiv.org/abs/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](https://openreview.net/forum?id=SkeHuCVFDz), 2020. URL <https://openreview.net/forum?id=SkeHuCVFDz>.
- Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. [arXiv preprint arXiv:2508.01191](https://arxiv.org/abs/2508.01191), 2025.
- 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 APPENDIX

A.1 MEDCALC BENCH DATA DETAILS

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 A1 shows that the trend of best to worst model performance remains the same across default and across empty variations.

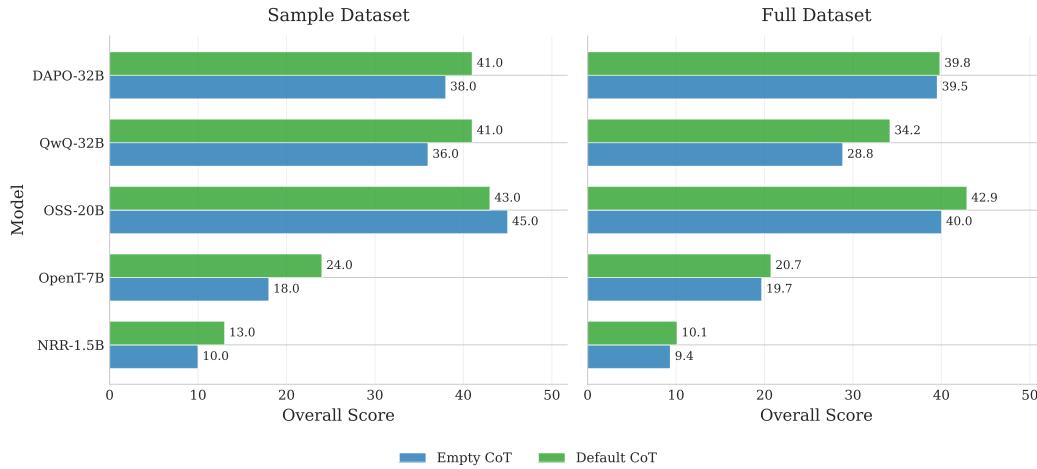


Figure A1: Model performance across all sampled and full data. The trends remain the same. Hence, the sample collected is a representative sample.

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 evaluated this model’s CoT in low reasoning level effort.

A.2 DETAILS ON REMOVING ANSWERS FROM EXPLANATIONS

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: Prompt for removing answer from an explanation.

A.3 INSTRUCTION INDUCTION ADDED DATA DETAILS

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

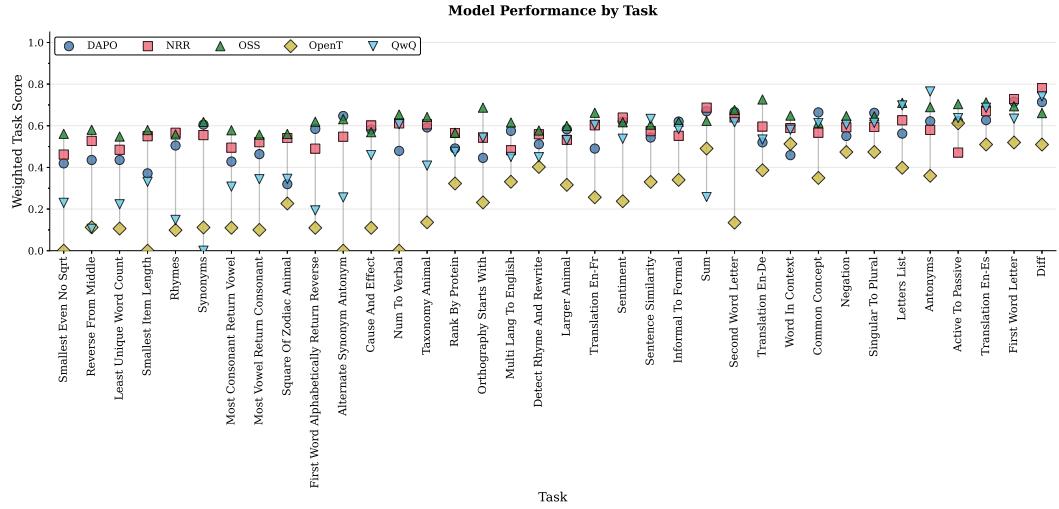


Figure A2: Model performance across all sampled data, including both the original instruction induction and newly introduced tasks.

A.4 DETAILS ON ENSEMBLED COT

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 A3,

we look at the distribution of candidates chosen from different pairs of generator models in various combinations of evaluated ensembles.

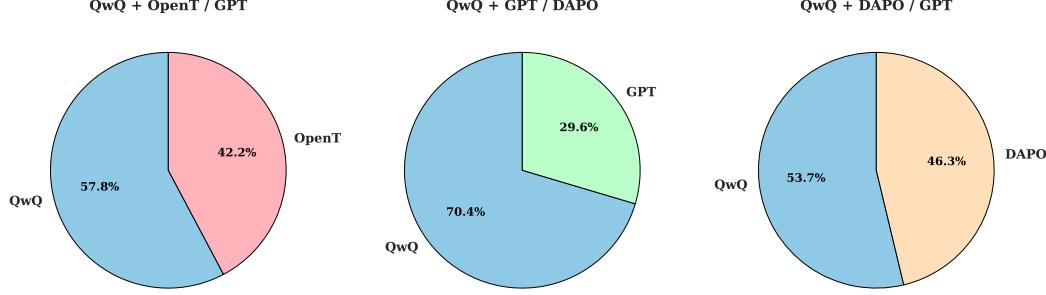


Figure A3: Proportion of selected candidate sentences from different generator models used to construct ensemble chain-of-thoughts across settings in MedCalc-Bench.

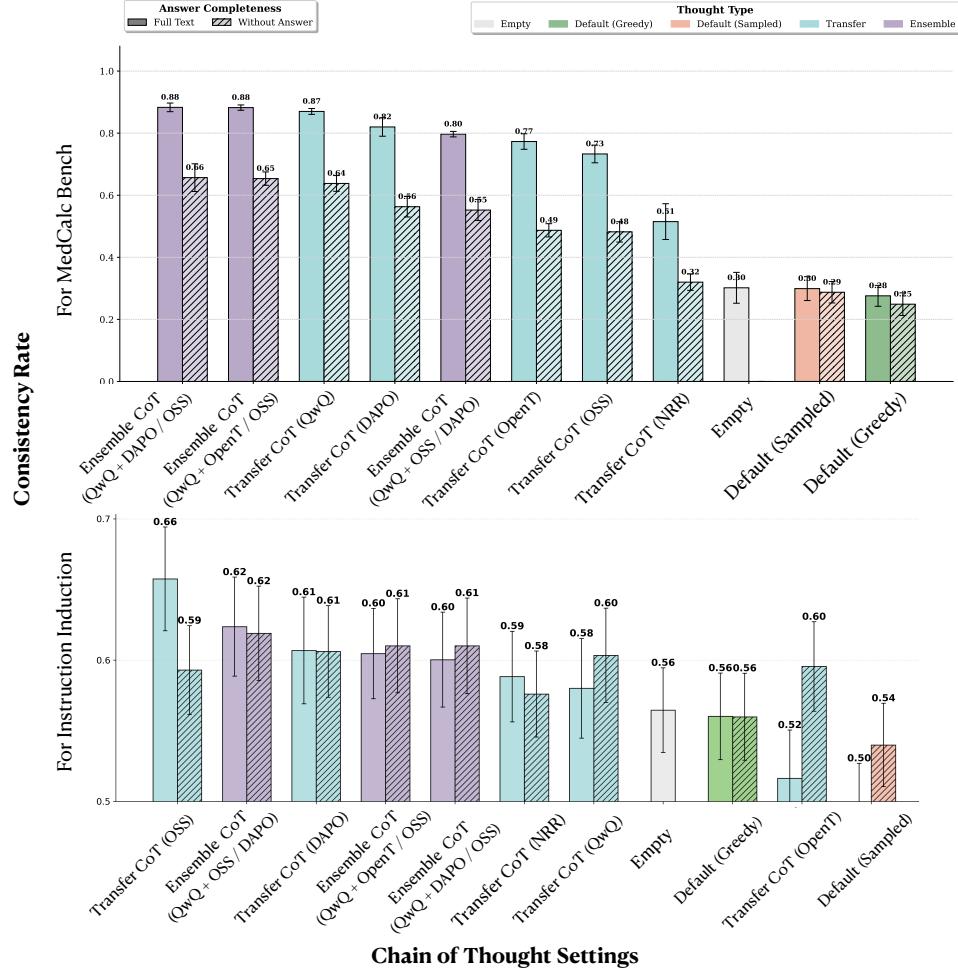


Figure A4: Average pairwise consistency across thought settings in MedCalc-Bench (above) and Instruction Induction (below). For thought variations indicating Ensemble CoT, models listed before the slash (/) serve as generators, while the model after the slash acts as the judge/evaluator. Results are reported both for the full text and for text with the final answer removed.

A.5 ADDITIONAL RESULTS

Figure A4 and Table A1 show consistency and accuracy for all settings. “Full text” denotes complete model-generated CoTs, while “w/o answer” denotes CoTs with direct answers removed, retaining only reasoning steps.

Table A1: 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 (Greedy) 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 |
| Default (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 |

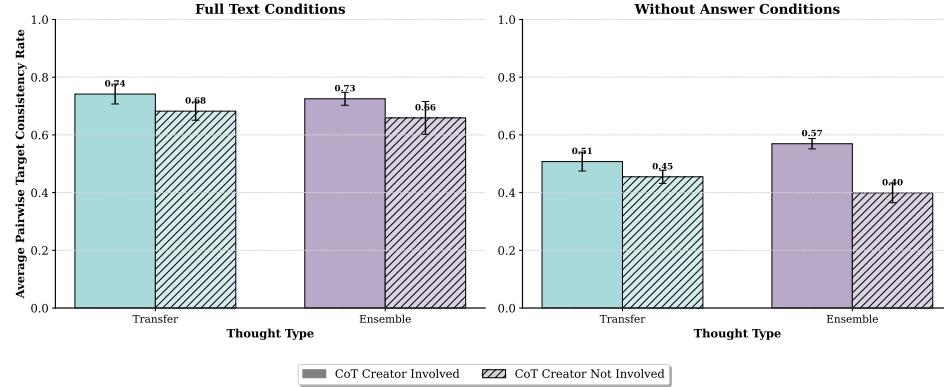


Figure A5: **LRMs transfer better using their own CoT than the CoT from other models.** Plots show the average pairwise consistency for transfer and ensemble thoughts, comparing MedCalc Bench cases where a model is involved as a creator/source of the CoT versus cases where none of the models tested were part of the source.

Fig. A5 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 CoT. Fig. A6 illustrates 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.

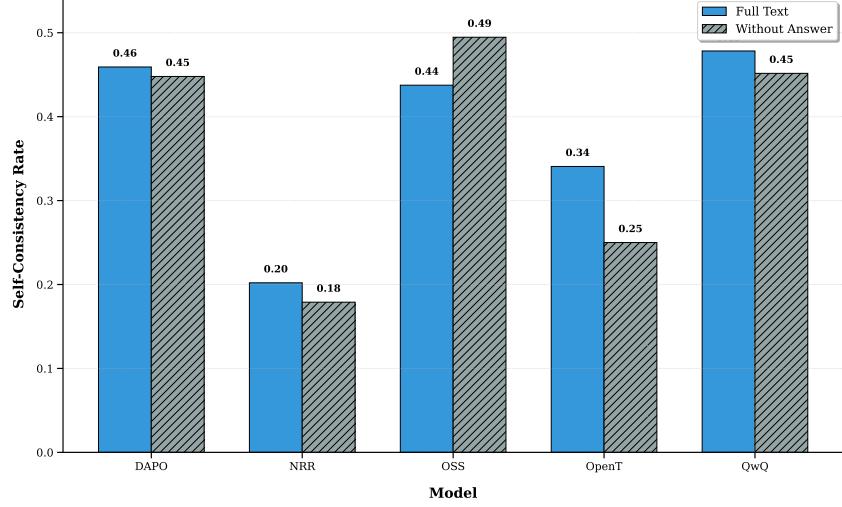


Figure A6: Model self-consistency rates for CoT generation on MedCalc-Bench. Bars compare answer consistency between default (greedy) and sampling-based decoding strategies, with results shown separately for full CoT text (left bar) and CoT with explicit answers removed (right bar).

A.6 VALIDATION WITH MEDCALC-BENCH-VERIFIED

Following completion of our main analysis, Khandekar et al. (2024) released MedCalc-Bench-Verified, an updated version with refined ground truth annotations. To validate the robustness of our findings, we re-evaluated our models on the overlapping subset of test samples using the corrected annotations. Our consistency trends remained substantively unchanged, confirming that our results are not artifacts of the original benchmark’s annotation quality.

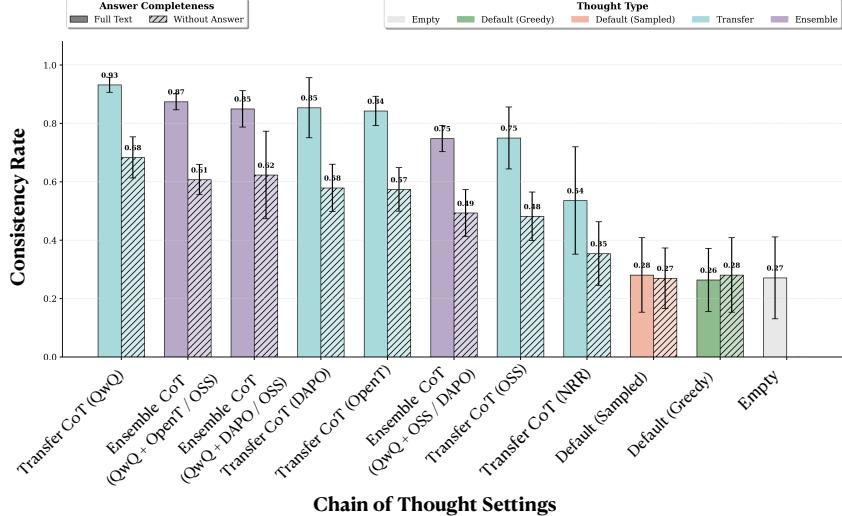


Figure A7: Average pairwise consistency across thought settings in MedCalc-Bench-Verified. For thought variations indicating Ensemble CoT, models listed before the slash (/) serve as generators, while the model after the slash acts as the judge/evaluator. Results are reported both for the full text and for text with the final answer removed.

A.7 USER STUDY DETAILS

While conducting the user study, no demographic or any other personal identifying information was collected.

Fig. A8 shares a detailed overview of user study scores. We also share the detailed results of Wilcoxon signed-rank test and paired t-test in Table A2.

Table A2: Pairwise Significance Matrix: Mean Differences (Wilcoxon Test)

| | OSS | DAPO | QwQ+DAPO/OSS | QwQ+OSS/DAPO |
|--------------------------|-----|-----------|--------------|--------------|
| <i>Clarity of Steps</i> | | | | |
| OSS | - | -1.16 *** | -1.58 *** | -0.31 *** |
| DAPO | - | - | -0.41 *** | +0.85 *** |
| QwQ+DAPO/OSS | | | - | +1.26 *** |
| QwQ+OSS/DAPO | | | | - |
| <i>Ease of Following</i> | | | | |
| OSS | - | -0.96 *** | -1.37 *** | -0.12 |
| DAPO | - | - | -0.41 *** | +0.84 *** |
| QwQ+DAPO/OSS | | | - | +1.25 *** |
| QwQ+OSS/DAPO | | | | - |
| <i>Confidence</i> | | | | |
| OSS | - | -1.05 *** | -1.36 *** | -0.15* |
| DAPO | - | - | -0.31 *** | +0.91 *** |
| QwQ+DAPO/OSS | | | - | +1.21 *** |
| QwQ+OSS/DAPO | | | | - |
| <i>Best Overall</i> | | | | |
| 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.

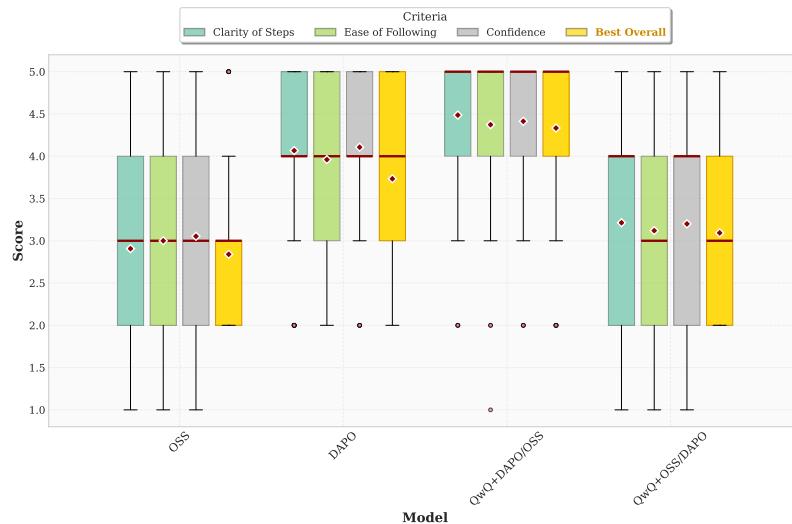


Figure A8: **User study results across evaluation criteria.** Box plots show CoT evaluation scores for each model across four criteria: Clarity of Steps, Ease of Following, Confidence, and Best Overall ranking. Red lines indicate medians, diamonds indicate means. Higher scores are better.