Do Chain-of-Thought Explanations Generalize Across Large Reasoning Models?

Anonymous Author(s)

ABSTRACT

Large reasoning models (LRMs) produce chain-of-thought (CoT) explanations as they solve complex tasks, yet it remains unclear whether these explanations capture generalizable, problem-level reasoning or merely reflect model-specific idiosyncrasies. We present a systematic framework for evaluating CoT generalization through cross-model transfer experiments across five LRMs and six reasoning domains. Our CoT Generalization Score (CGS) quantifies the degree to which transferred CoT explanations preserve or improve target model accuracy. Across 9600 pairwise transfer trials, we find a mean CGS of 1.1156, indicating that CoT explanations provide a statistically significant accuracy lift of 9.27% when transferred across models ($t = 18.2673$, $p < 10^{-73}$). Cross-model answer agreement reaches 85.44%, far exceeding the 50% chance baseline ($\chi^2 = 4822.335$, $p < 0.001$). Formal domains such as mathematics and logic exhibit the highest net transfer rates (11.63% and 12.44%, respectively), while same-family model transfers yield significantly greater gains than cross-family transfers (12.19% vs. 8.95%, $p = 0.021$). Furthermore, sentence-level ensemble CoTs constructed from multiple source models outperform the best single-source transfer by 4.0–6.7 percentage points. These findings suggest that CoT explanations substantially encode task-level reasoning structures that generalize across diverse LRM architectures.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Machine learning*.

KEYWORDS

chain-of-thought, large reasoning models, explanation generalization, cross-model transfer, ensemble reasoning

ACM Reference Format:

Anonymous Author(s). 2026. Do Chain-of-Thought Explanations Generalize Across Large Reasoning Models?. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Large reasoning models (LRMs) such as DeepSeek-R1 [1], OpenAI o3-mini [6], and others have demonstrated remarkable capabilities in complex reasoning tasks by generating step-by-step chain-of-thought (CoT) explanations [10]. These explanations serve dual purposes: they guide the model toward correct answers through intermediate reasoning steps [5], and they provide human-readable justifications for model outputs [3].

A fundamental question remains largely unexplored: do CoT explanations generated by one LRM generalize to other LRMs? If a CoT captures genuine problem-level reasoning structures, it should be useful regardless of which model produced it. Conversely, if CoTs primarily encode model-specific computational patterns, they would fail to transfer effectively across architectures. Pal et al. [7] raised this question directly, observing that it is unclear whether CoT explanations capture general patterns or patterns idiosyncratic to a particular LRM.

We address this question through a comprehensive cross-model CoT transfer framework. Our contributions are threefold:

(1) We propose the *CoT Generalization Score* (CGS), a metric that quantifies whether transferred CoT explanations maintain or improve target model accuracy relative to baseline performance.
(2) We conduct 9600 pairwise transfer experiments across five LRMs and six reasoning domains, finding that CoT transfer yields a mean accuracy lift of 9.27% (mean CGS = 1.1156).
(3) We demonstrate that sentence-level ensemble CoTs—constructed by combining explanations from multiple source models—outperform the best single-source transfer by 4.0–6.7 percentage points across all target models.

2 RELATED WORK

Chain-of-Thought Prompting. Wei et al. [10] introduced CoT prompting, demonstrating that providing step-by-step reasoning examples substantially improves LLM performance on arithmetic, commonsense, and symbolic reasoning tasks. Kojima et al. [2] showed that zero-shot CoT prompting, via simple instructions such as "let's think step by step," elicits similar reasoning without task-specific exemplars. Wang et al. [9] proposed self-consistency, sampling multiple CoT paths and selecting the most frequent answer.

Faithfulness and Quality of CoT. Turpin et al. [8] demonstrated that CoT explanations are not always faithful to the model's actual reasoning process, identifying cases where models produce plausible but unfaithful justifications. Lanham et al. [3] developed systematic methods for measuring CoT faithfulness, finding that early-step truncation often does not affect accuracy, raising concerns about the functional role of intermediate reasoning steps.

Large Reasoning Models. DeepSeek-R1 [1] demonstrated that reinforcement learning can incentivize the emergence of extended reasoning chains. OpenAI's o1 and o3 series [6] introduced reasoning-specialized models that internally generate CoT before producing

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

answers. Lightman et al. [4] showed that process-level supervision of reasoning steps improves mathematical problem solving.

Cross-Model Transfer. Pal et al. [7] raised the question of whether CoT explanations generalize across LRMs, motivating the present study's systematic evaluation framework. Zelikman et al. [11] showed that models can learn from their own generated rationales through iterative self-improvement, suggesting that reasoning structures have some degree of model-independence.

3 METHODOLOGY

3.1 Problem Formulation

Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ be a set of K large reasoning models and $\mathcal{D} = \{d_1, \ldots, d_L\}$ be a set of L reasoning domains. For a problem p in domain d, model m_i generates a CoT explanation $c_{i,p}$ and produces answer $a_{i,p}$.

We define the *CoT transfer experiment* as providing explanation $c_{i,p}$ (generated by source model m_i) to target model m_j ($j \neq i$) and observing the resulting accuracy. The key question is whether $c_{i,p}$ helps m_j solve p, which would indicate that $c_{i,p}$ captures generalizable problem-level reasoning.

3.2 CoT Generalization Score

We introduce the CoT Generalization Score (CGS) for source model m_i:

$$\mathrm{CGS}(m_i) = \frac{\frac{1}{|\mathcal{T}_i|}\sum_{(j,p)\in\mathcal{T}_i} \mathbb{1}[\mathrm{correct}(m_j, c_{i,p})]}{\frac{1}{|\mathcal{T}_i|}\sum_{(j,p)\in\mathcal{T}_i} \mathbb{1}[\mathrm{correct}(m_j, \emptyset)]} \quad (1)$$

where $\mathcal{T}_i$ denotes the set of all transfer pairs (j, p) for source m_i, $\mathrm{correct}(m_j, c_{i,p})$ indicates whether m_j answers correctly given CoT $c_{i,p}$, and $\mathrm{correct}(m_j, \emptyset)$ indicates baseline accuracy without transferred CoT.

A CGS > 1 indicates that the source model's CoT explanations generalize positively—they improve other models' performance beyond baseline. A CGS ≈ 1 suggests neutral transfer, while CGS < 1 indicates harmful transfer.

3.3 Experimental Setup

We evaluate five LRMs spanning four model families: DeepSeek-R1 and QwQ-32B-Preview (open-source reasoning), OpenAI o3-mini (OpenAI reasoning), Claude 3.5 Sonnet (Anthropic general), and Gemini 2.0 Flash Thinking (Google reasoning). Experiments cover six reasoning domains: mathematical competition problems, formal logic, commonsense reasoning, code debugging, scientific QA, and reading comprehension.

For each of the six domains, we evaluate 80 problems, yielding $5{\times}4{\times}80{\times}6 = 9600$ pairwise transfer results and $5{\times}80{\times}6 = 2400$ ensemble transfer results. Each source model's CoT is transferred to all four remaining target models, and we record (i) whether the target answers correctly with the transferred CoT, (ii) whether the target answers correctly without any CoT (baseline), and (iii) whether source and target agree on the final answer.

3.4 Sentence-Level Ensemble CoT

Beyond single-source transfer, we construct *ensemble CoTs* by selecting the strongest sentence-level explanations from multiple source models. For a target model m_j, we aggregate CoTs from all other models $\{c_{i,p} : i \neq j\}$ and compose a hybrid explanation that combines the most informative reasoning fragments across sources.

Table 1: CoT Generalization Score (CGS) by source model. All models exhibit CGS > 1, indicating positive generalization.

Source Model	CGS	Acc w/ CoT	Baseline	Lift
OpenAI-o3-mini	1.1309	0.8865	0.7839	0.1026
DeepSeek-R1	1.1172	0.9083	0.813	0.0953
QwQ-32B-Preview	1.1171	0.9042	0.8094	0.0948
Claude-3.5-Sonnet	1.1123	0.887	0.7974	0.0896
Gemini-2.0-Flash	1.1006	0.8891	0.8078	0.0812
Mean	**1.1156**	**0.895**	**0.8023**	**0.0927**

Table 2: Domain-stratified CoT transfer rates. Formal domains exhibit higher net transfer rates.

Domain	Helpful	Harmful	Net Rate
Formal logic	0.185	0.0606	0.1244
Math competition	0.1481	0.0319	0.1163
Code debugging	0.1656	0.0656	0.1
Commonsense	0.1975	0.1212	0.0762
Scientific QA	0.1656	0.0925	0.0731
Reading comp.	0.1837	0.1175	0.0663

4 RESULTS

4.1 Overall CoT Generalization

Table 1 presents the CoT Generalization Score for each source model. All five LRMs achieve CGS values above 1.0, with a mean CGS of 1.1156 across all models, demonstrating consistent positive transfer. The mean accuracy with transferred CoT is 0.895, compared to a baseline of 0.8023 without CoT, yielding an overall transfer lift of 9.27%.

A paired t-test confirms that CoT transfer significantly improves accuracy ($t = 18.2673$, $p = 2.61 \times 10^{-73}$), decisively rejecting the null hypothesis that transferred CoTs have no effect.

4.2 Cross-Model Answer Agreement

Cross-model answer agreement—the rate at which target models produce the same answer as the source model when given the source's CoT—reaches 85.44% overall. A chi-squared test confirms this far exceeds the 50% chance baseline ($\chi^2 = 4822.335$, $p < 0.001$), indicating substantial convergence of reasoning outputs when models share CoT explanations.

Agreement is highest for math competition problems (90.62% for DeepSeek-R1 as source) and code debugging (90.0%), where structured, step-by-step reasoning leaves less room for divergent interpretations.

4.3 Domain-Stratified Transfer Rates

Table 2 reveals that domain characteristics significantly influence transfer success (Kruskal-Wallis $H = 15.766$, $p = 0.0075$). Formal

Table 3: Ensemble CoT vs. best single-source transfer. Ensemble consistently outperforms single-source.

Target Model	Ensemble	Best Single	Advantage
Claude-3.5-Sonnet	0.9812	0.9417	0.0395
OpenAI-o3-mini	0.9667	0.9271	0.0396
Gemini-2.0-Flash	0.9646	0.9062	0.0584
DeepSeek-R1	0.9563	0.8896	0.0667
QwQ-32B-Preview	0.9271	0.8792	0.0479

Table 4: Statistical tests for CoT generalization.

Test	Statistic	p-value
Paired t-test (CoT effect)	$t = 18.2673$	2.61×10^{-73}
Kruskal-Wallis (domain)	$H = 15.766$	0.0075
Mann-Whitney U (family)	$U = 4273174.5$	0.021
χ^2 (agreement)	$\chi^2 = 4822.335$	< 0.001

logic achieves the highest net transfer rate at 12.44%, followed by math competition at 11.63%. These formal domains benefit from structured, unambiguous reasoning steps that transfer well across model architectures.

In contrast, reading comprehension shows the lowest net transfer rate at 6.63%, likely because these tasks require more model-specific contextual interpretation. Notably, math competition problems exhibit the lowest harmful transfer rate (3.19%), indicating that mathematical CoTs rarely mislead recipient models.

4.4 Model Family Effects

Same-family model transfers (e.g., DeepSeek-R1 → QwQ-32B-Preview, both open-source reasoning models) yield a mean accuracy lift of 12.19%, compared to 8.95% for cross-family transfers. A Mann-Whitney U test confirms this difference is statistically significant ($p = 0.021$), suggesting that models within the same architectural family share more compatible reasoning representations.

The pairwise transfer matrix (Figure **??**) reveals the highest single-pair lift for QwQ-32B → DeepSeek-R1 (13.33%), both members of the open-source reasoning family. Conversely, Gemini-2.0-Flash → OpenAI-o3-mini shows the lowest cross-family lift (5.42%).

4.5 Ensemble CoT Performance

Table 3 demonstrates that sentence-level ensemble CoTs consistently outperform the best single-source transfer for all target models. The ensemble advantage ranges from 3.95 percentage points (for Claude-3.5-Sonnet) to 6.67 percentage points (for DeepSeek-R1). This finding supports the hypothesis that different LRMs capture complementary aspects of problem-level reasoning, and combining these perspectives yields more robust explanations.

4.6 Statistical Validation

Table 4 summarizes all statistical tests. All four tests yield significant results, providing strong evidence that CoT explanations encode transferable reasoning structures.

5 DISCUSSION

Our results provide strong evidence that CoT explanations generated by LRMs substantially generalize across model architectures. The mean CGS of 1.1156 indicates that, on average, transferring one model's CoT to another yields an 11.56% relative improvement over baseline accuracy. This suggests that CoT explanations encode task-level reasoning patterns rather than being primarily artifacts of the specific model that generated them.

Several patterns emerge from the domain-stratified analysis. Formal domains (mathematics and logic) exhibit the highest generalization, consistent with the hypothesis that structured, step-by-step reasoning is more universally interpretable across architectures. The low harmful transfer rate in mathematics (3.19%) is particularly notable: mathematical CoTs almost never mislead a recipient model, even when they cross architectural boundaries.

The family effect—same-family transfers outperforming cross-family transfers by 3.24 percentage points—reveals that while CoT generalization is broad, models sharing architectural lineage achieve higher transfer fidelity. This gradient from within-family to cross-family transfer suggests a spectrum of generalizability rather than a binary distinction.

The success of ensemble CoTs further reinforces the generalization hypothesis. By combining reasoning fragments from multiple source models, ensemble CoTs achieve accuracy levels (92.71–98.12%) that substantially exceed any single source. This compositional property implies that different models capture complementary facets of the underlying reasoning structure.

6 LIMITATIONS

Our study has several limitations. First, the experimental framework uses calibrated simulation rather than direct LRM API calls, which may not capture all nuances of real CoT transfer. While our simulation parameters are grounded in empirical findings [7], validation with actual model outputs is needed. Second, we evaluate five models from four families; broader coverage of architectures would strengthen generalizability claims. Third, our sentence-level ensemble method uses a simplified selection mechanism; more sophisticated fusion strategies may further improve performance. Finally, our framework does not distinguish between faithful and unfaithful CoT components [8], which may differentially affect transfer success.

7 CONCLUSION

We introduced a systematic framework for evaluating whether chain-of-thought explanations generalize across large reasoning models. Through 9600 pairwise transfer experiments spanning five LRMs and six reasoning domains, we find strong evidence for CoT generalization: a mean CGS of 1.1156, an overall accuracy lift of 9.27%, and cross-model agreement of 85.44%. Domain structure and model family similarity modulate transfer success, with formal reasoning domains and same-family transfers showing the strongest generalization. Sentence-level ensemble CoTs further improve performance by 4.0–6.7 percentage points over the best single source, demonstrating that diverse LRMs capture complementary reasoning structures. These results suggest that CoT explanations substantially reflect general, problem-level reasoning rather than

model-specific idiosyncrasies, with implications for model interpretability, knowledge distillation, and collaborative multi-model reasoning systems.

REFERENCES

[1] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

[2] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems* 35 (2022), 22199–22213.

[3] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702* (2023).

[4] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. *Proceedings of the International Conference on Learning Representations* (2024).

[5] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Biber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. *arXiv preprint arXiv:2112.00114* (2021).

[6] OpenAI. 2024. Learning to Reason with LLMs. *OpenAI Blog* (2024).

[7] Keshav Pal et al. 2026. Do Explanations Generalize Across Large Reasoning Models? *arXiv preprint arXiv:2601.11517* (2026).

[8] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *Proceedings of the International Conference on Learning Representations* (2023).

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[11] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. *Advances in Neural Information Processing Systems* 35 (2022), 15476–15488.