Chain of Thought as a Cross-Model Homogenizer: CoT Increases Convergence Between LLMs Without Reducing Individual Diversity

Ari Holtzman and Idea-Explorer

Abstract

Large language models are increasingly converging in their outputs—different models produce remarkably similar responses to the same prompts. Chain of Thought (CoT) prompting is now ubiquitous in modern LLMs, yet its specific effect on this convergence phenomenon has never been directly measured. We conduct the first empirical study of CoT's effect on both within-model and cross-model output convergence, prompting four LLMs from different families (GPT-4.1, CLAUDE SONNET 4.5, GEMINI 2.5 FLASH, LLAMA 3.1 70B) with and without CoT across 46 questions spanning reasoning, creative, and opinion tasks. We find that CoT dramatically increases cross-model convergence (Cohen's $d = 0.93$, $p < 0.00001$), raising average pairwise cosine similarity from 0.634 to 0.832. However, CoT does *not* increase within-model convergence ($d = -0.074$), meaning individual models maintain similar diversity across repeated samples. This asymmetry reveals that CoT acts as a *cross-model homogenizer*: it funnels different architectures toward shared reasoning templates while preserving each model's individual variability. The effect is strongest for reasoning tasks ($d = 1.19$) and persists across creative ($d = 0.76$) and opinion ($d = 0.90$) tasks. These findings have direct implications for AI deployment strategies that rely on model diversity, benchmark design, and the growing concern over LLM output homogenization.

1 Introduction

Large language models are converging. Despite differences in architecture, training data, and alignment procedures, models from OpenAI, Anthropic, Google, and Meta produce increasingly similar outputs to the same prompts [Wenger and Kenett, 2025, Mirka et al., 2025]. This homogenization threatens the value of maintaining multiple AI systems—if all models give essentially the same answer, the benefits of model diversity for creative applications, cultural pluralism, and robust decision-making are diminished.

At the same time, Chain of Thought (CoT) prompting has become ubiquitous. Since Wei et al. [2022] showed that prompting models to "think step by step" dramatically improves reasoning, CoT has been adopted across virtually every LLM application and has been integrated into model training itself [Chen et al., 2025]. Yet while CoT's effect on *accuracy* is well-studied, its effect on output *diversity* across models is not. Recent work provides suggestive evidence: Xu and Zhang [2026] showed that English-language CoT creates "convergence basins" in thinking space within individual models, and Jain et al. [2025] demonstrated that reasoning tasks achieve only 2–3 distinct solution strategies even with diversity-promoting techniques. However, no study has directly measured whether CoT increases or decreases the similarity of outputs *across different model families*.

We conduct the first direct empirical test. We prompt four LLMs from different families—GPT-4.1 (OpenAI), CLAUDE SONNET 4.5 (Anthropic), GEMINI 2.5 FLASH (Google), and LLAMA 3.1 70B (Meta)—with and without CoT on 46 questions spanning three task types: reasoning (BIG-

Bench Hard), creative writing (NoveltyBench), and opinion expression (World Values Survey). For each model–condition–question combination, we generate three responses at temperature 0.7 and measure convergence using pairwise cosine similarity of sentence embeddings at two levels: within-model (how similar are repeated samples from the same model?) and cross-model (how similar are responses from different models?).

Our results reveal a striking asymmetry. CoT increases cross-model convergence by 31% in relative terms (pairwise cosine similarity from 0.634 to 0.832), with a large effect size (Cohen's $d = 0.93$, $p < 0.00001$). But CoT does *not* increase within-model convergence ($d = -0.074$). In other words, CoT makes different models sound alike without making any individual model more repetitive. The effect is strongest for reasoning tasks ($d = 1.19$) and persists across creative ($d = 0.76$) and opinion tasks ($d = 0.90$). We also find a surprising dissociation: CoT *decreases* answer agreement on reasoning tasks (from 44.8% to 22.9%) despite dramatically increasing semantic similarity—models produce similar-looking reasoning chains that sometimes lead to different conclusions.

These findings have direct implications for AI deployment, benchmark design, and the ongoing discussion around LLM homogenization. Organizations that maintain multiple model providers for diversity should be aware that CoT significantly reduces the effective diversity of their model ensemble. Benchmarks that require CoT may underestimate true inter-model differences. And the convergence phenomenon itself appears to be driven, at least in part, by a specific and modifiable prompting practice rather than being an inevitable consequence of scale.

We make the following contributions:

- We provide the first direct empirical evidence that CoT increases cross-model convergence (Cohen's $d = 0.93$) while not increasing within-model convergence, identifying CoT as a cross-model homogenizer.
- We demonstrate that this effect holds across reasoning, creative, and opinion tasks, with the strongest effect on reasoning ($d = 1.19$), and identify a dissociation between semantic similarity and answer agreement under CoT.
- We discuss implications for model diversity strategies and benchmark design, and identify response length and shared training data as potential mechanisms warranting further investigation.

2 Related Work

Chain of Thought prompting. Wei et al. [2022] introduced CoT prompting by demonstrating that including intermediate reasoning steps in few-shot exemplars unlocks reasoning abilities in large language models. This approach has been extended to zero-shot settings [Kojima et al., 2022], self-consistency decoding [Wang et al., 2023], and has been incorporated directly into model training via outcome-based reinforcement learning [Chen et al., 2025]. While the accuracy benefits of CoT are well-established, its effects on output diversity and cross-model similarity have received far less attention. Our work addresses this gap by directly measuring how CoT changes the similarity structure of outputs across model families.

LLM output homogenization. A growing body of work documents that LLM outputs are converging across model families. Wenger and Kenett [2025] found that 22 LLMs from different families show dramatically lower creative variability than humans (effect sizes 1.4–2.2), even when controlling for response structure. Mirka et al. [2025] provided a comprehensive review of LLM-driven content homogenization across research ideation, essay writing, survey responses, and creative tasks. Jain et al. [2025] showed that the degree of homogenization is task-dependent, with problem-solving tasks achieving only 2–3 distinct strategies out of 5 generations even with diversity-promoting techniques. Unlike these studies, which document homogenization as a phenomenon, we test whether a specific prompting technique—CoT—is a causal contributor.

Language of thought and diversity. Most closely related to our work, Xu and Zhang [2026] demonstrated that standard English-language CoT creates "convergence basins" in thinking space within individual models. By controlling the thinking language via translated prefixes, they showed that non-English reasoning yields 5.3–7.7 point improvements in output diversity, and that languages geometrically farther from English in representation space produce greater diversity. Their study focuses on diversity within a single model across different thinking languages. We complement

their findings by measuring convergence *across* model families under the same language (English), and find that CoT's homogenizing effect extends to the cross-model setting.

Faithfulness of CoT reasoning. Several studies have shown that CoT explanations can be systematically unfaithful. Turpin et al. [2023] demonstrated that models produce confident CoT explanations that never mention biasing features, even when those features determine the answer. Chen et al. [2025] extended this to reasoning models (Claude 3.7, DeepSeek R1), finding faithfulness rates of only 25–39%. Afzal et al. [2025] showed that models encode information about CoT success *before* generating any reasoning tokens, suggesting that CoT may be post-hoc rationalization rather than genuine computation. These findings suggest a mechanism for our observed convergence: if CoT is partially a surface-level rationalization, models may converge on shared rationalization templates drawn from similar training data, even when their underlying computations differ.

3 Methodology

We design a controlled experiment to isolate the effect of CoT prompting on output convergence. The core idea is simple: prompt the same set of questions to multiple LLMs with and without CoT, then measure how similar the resulting outputs are both within and across models.

3.1 Models

We select four models from different families to ensure that any observed convergence reflects cross-architecture patterns rather than within-family similarity:

- GPT-4.1 (OpenAI) — accessed via the OpenAI API
- CLAUDE SONNET 4.5 (Anthropic) — accessed via OpenRouter
- GEMINI 2.5 FLASH (Google) — accessed via OpenRouter
- LLAMA 3.1 70B (Meta) — accessed via OpenRouter

These models span the four major LLM families, differ in architecture and training procedures, and represent a mix of proprietary and open-weight systems.

3.2 Datasets

We sample questions from three established benchmarks spanning different task types to test whether CoT's effect on convergence varies with the nature of the task:

Dataset	Task Type	Questions	Source
BBH	Reasoning	16	Suzgun et al. [2023]
NOVELTYBENCH	Creative	15	Xu and Zhang [2026]
WVS	Opinion	15	Xu and Zhang [2026]

Table 1: Datasets used in our study. We sample 46 total questions across three task types to cover reasoning, creative, and opinion domains.

Reasoning (BBH). We sample 16 questions from BIG-Bench Hard [Suzgun et al., 2023], drawing two questions each from eight diverse tasks: date understanding, causal judgement, disambiguation QA, logical deduction, navigation, reasoning about colored objects, sports understanding, and web of lies. These tasks have ground-truth answers, allowing us to also measure answer agreement.

Creative (NOVELTYBENCH). We use 15 questions from NoveltyBench [Xu and Zhang, 2026], which contains open-ended creative prompts such as "Write a short love poem with 4 lines." These tasks have no single correct answer, and diversity is inherently valuable.

Opinion (WVS). We use 15 questions from the World Values Survey [Xu and Zhang, 2026], which ask about values and preferences (e.g., "How important is family in your life?") with fixed response options. These test whether CoT homogenizes value expression across models.

3.3 Experimental Design

Conditions. We compare two prompting conditions:

- **Direct:** System prompt instructs the model to "Answer the question directly and concisely. Do not explain your reasoning." User prompt ends with "Answer directly and concisely."
- **CoT:** System prompt instructs the model to "Answer the question by thinking step by step. Show your reasoning before giving your final answer." User prompt ends with "Let's think step by step."

Sampling. For each (model, condition, question) triple, we generate 3 responses at temperature 0.7 with a maximum of 400 tokens. Temperature 0.7 is standard for diversity measurement—low enough to produce coherent outputs, high enough to reveal meaningful variation. This yields $46 \times 4 \times 2 \times 3 = 1{,}104$ total API calls, all of which completed successfully.

3.4 Metrics

Pairwise Cosine Similarity (PCS). Our primary metric. We embed all responses using ALL-MINILM-L6-V2 [Reimers and Gurevych, 2019], a widely used sentence embedding model that produces 384-dimensional L2-normalized vectors. For a set of n responses, we compute the average cosine similarity over all $\binom{n}{2}$ pairs. Higher PCS indicates greater convergence.

We compute PCS at two levels:

- **Within-model PCS:** Average pairwise similarity among the 3 responses from the same model, same condition, same question. Measures how repetitive a single model is.
- **Cross-model PCS:** Average pairwise similarity among responses from *different* models, same condition, same question. Measures how similar different models are to each other.

Answer Agreement Rate (AAR). For BBH reasoning tasks with ground-truth answers, we measure the fraction of cross-model response pairs that produce the same final answer.

Lexical diversity. We compute Distinct-1 (ratio of unique unigrams to total unigrams) and Distinct-2 (ratio of unique bigrams to total bigrams) to measure surface-level lexical variation.

3.5 Statistical Analysis

We use the Wilcoxon signed-rank test for all paired comparisons between conditions. This non-parametric test is appropriate because PCS values are bounded and may not be normally distributed. We report Cohen's d as our primary effect size measure, using the conventional thresholds of 0.2 (small), 0.5 (medium), and 0.8 (large). All tests use a significance threshold of $\alpha = 0.05$. Each question serves as a paired observation—we compare the PCS for that question under direct prompting versus CoT prompting.

4 Results

4.1 Cross-Model Convergence: CoT Dramatically Increases Similarity

Our main finding is that CoT substantially increases the similarity of outputs across different model families. Table 2 reports cross-model PCS for each task type under direct and CoT conditions.

Across all 46 questions, CoT raises average cross-model PCS from 0.634 to 0.832—a 31% relative increase. The effect is highly significant ($p < 0.00001$) with a large effect size ($d = 0.93$). Figure 1 illustrates this pattern: under direct prompting, cross-model similarity varies widely across questions, while under CoT, it is consistently high and tightly clustered.

The effect is strongest for reasoning tasks ($d = 1.19$), where CoT raises cross-model PCS from 0.654 to 0.886. This is consistent with the hypothesis that step-by-step reasoning activates shared algorithmic patterns learned from overlapping training data. Creative tasks show the smallest (but still large) effect ($d = 0.76$), while opinion tasks fall in between ($d = 0.90$).

Category	Direct	CoT	p-value	Cohen's d
Overall	0.634 $\pm$ 0.213	**0.832** $\pm$ 0.080	<0.00001	**+0.93**
Reasoning	0.654 $\pm$ 0.188	**0.886** $\pm$ 0.042	0.002	**+1.19**
Creative	0.652 $\pm$ 0.232	**0.824** $\pm$ 0.052	0.018	**+0.76**
Opinion	0.595 $\pm$ 0.212	**0.783** $\pm$ 0.097	0.008	**+0.90**

Table 2: Cross-model pairwise cosine similarity (PCS) under direct and CoT prompting. CoT increases cross-model similarity across all task types, with effect sizes ranging from large ($d = 0.76$) to very large ($d = 1.19$). Higher values indicate greater convergence. Best results (higher convergence) in **bold**.

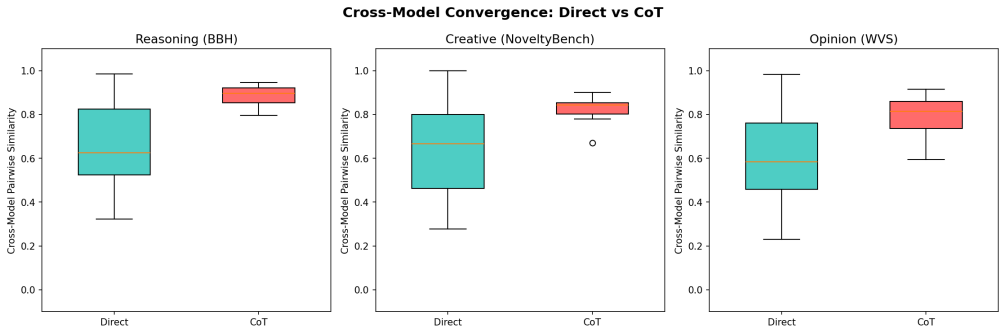

Figure 1: Cross-model pairwise cosine similarity under direct and CoT prompting, broken down by task type. CoT consistently increases cross-model convergence, with the tightest clustering for reasoning tasks. Box plots show median, interquartile range, and outliers.

4.2 Within-Model Convergence: CoT Does Not Increase Repetitiveness

In contrast to the dramatic cross-model effect, CoT has virtually no effect on within-model convergence. Table 3 shows that individual models produce responses of similar diversity with and without CoT.

Overall within-model PCS is nearly identical between conditions (0.871 vs. 0.857, $d = -0.074$). The small statistically significant p-value ($p = 0.003$) reflects the large sample size rather than a meaningful effect—the Cohen's d is negligible. For reasoning and creative tasks, the effect is essentially zero ($d = -0.004$ and $d = +0.028$, respectively). The only notable within-model change is for opinion tasks, where CoT actually *reduces* within-model similarity ($d = -0.27$, $p = 0.013$), suggesting that step-by-step reasoning causes models to explore different perspectives across samples.

Figure 2 confirms this pattern visually. The within-model distributions are largely overlapping between conditions, in stark contrast to the cross-model distributions in figure 1.

4.3 The Asymmetry: Cross-Model vs. Within-Model Effects

Figure 3 summarizes the effect sizes across all conditions, highlighting the core asymmetry. Cross-model Cohen's d values are large and positive (0.76–1.19), while within-model values hover near zero (-0.27 to +0.03). This asymmetry is the central finding of our study: CoT is a *cross-model* homogenizer but not a *within-model* homogenizer.

4.4 Answer Agreement vs. Semantic Similarity

For the BBH reasoning tasks that have ground-truth answers, we observe a surprising dissociation. CoT *decreases* answer agreement across models (from 44.8% to 22.9%, $p = 0.029$) despite dramatically increasing semantic similarity. Models under CoT produce reasoning chains that look very similar in structure, vocabulary, and approach—yet these chains sometimes lead to different final answers. This pattern is consistent with the literature on unfaithful CoT [Turpin et al., 2023]: the reasoning traces converge on shared templates while the underlying computation may diverge.

Category	Direct	CoT	p-value	Cohen's d
Overall	0.871 $\pm$ 0.195	0.857 $\pm$ 0.099	0.003	-0.074
Reasoning	0.909 $\pm$ 0.150	0.908 $\pm$ 0.043	0.083	-0.004
Creative	0.844 $\pm$ 0.230	0.851 $\pm$ 0.109	0.397	+0.028
Opinion	0.859 $\pm$ 0.193	0.807 $\pm$ 0.106	0.013	-0.268

Table 3: Within-model pairwise cosine similarity under direct and CoT prompting. CoT does not meaningfully change within-model similarity overall ($d = -0.074$). For opinion tasks, CoT actually *decreases* within-model similarity ($d = -0.27$, $p = 0.013$).

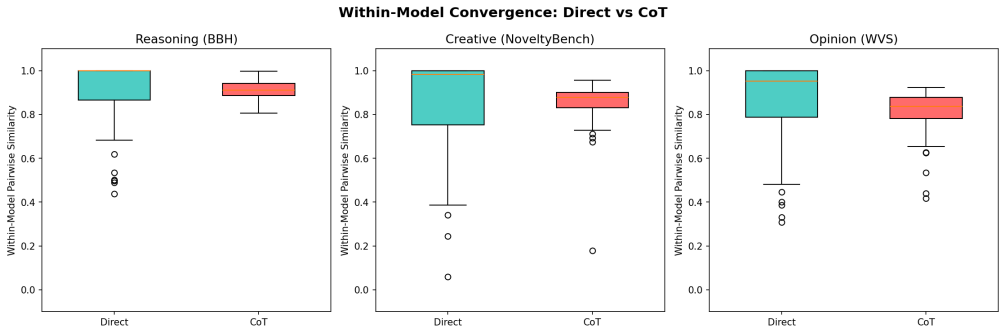

Figure 2: Within-model pairwise cosine similarity under direct and CoT prompting. Unlike the cross-model case (figure 1), within-model similarity remains stable across conditions, confirming that CoT does not make individual models more repetitive.

4.5 Lexical Diversity

CoT significantly reduces Distinct-1 (unique unigrams: 0.490 $\rightarrow$ 0.380, $p < 0.0001$) but increases Distinct-2 (unique bigrams: 0.348 $\rightarrow$ 0.681, $p < 0.0001$). The decrease in Distinct-1 reflects convergence on shared vocabulary (e.g., "Step 1,", "Let's consider"), while the increase in Distinct-2 reflects the greater length and elaboration of CoT responses, which naturally produce more unique bigram combinations.

4.6 Model-Level Patterns

Figure 4 shows the pairwise similarity between all model pairs under both conditions. Under direct prompting, model pairs show moderate and variable similarity (0.55–0.72). Under CoT, all model pairs converge to high similarity (0.78–0.90), with the spread between the most-similar and least-similar pairs shrinking substantially. No single model pair drives the effect—the convergence is consistent across all six pairwise comparisons.

5 Discussion

5.1 Why Does CoT Homogenize Across Models but Not Within?

The central puzzle of our results is the asymmetry: CoT increases cross-model similarity without increasing within-model similarity. We consider three potential explanations.

Shared reasoning templates. All four models were likely trained on overlapping corpora that contain similar step-by-step reasoning patterns—math textbooks, programming tutorials, educational materials. When prompted to reason step by step, models draw on these shared templates, producing structurally similar outputs. Within a single model, however, the stochastic sampling process (temperature 0.7) introduces variation that is orthogonal to the template structure, preserving within-model diversity.

Response length effects. CoT responses are 20–30$\times$ longer than direct responses (e.g., $\sim$190 vs. $\sim$12 words for reasoning tasks). Longer texts share more common phrases and structural elements,

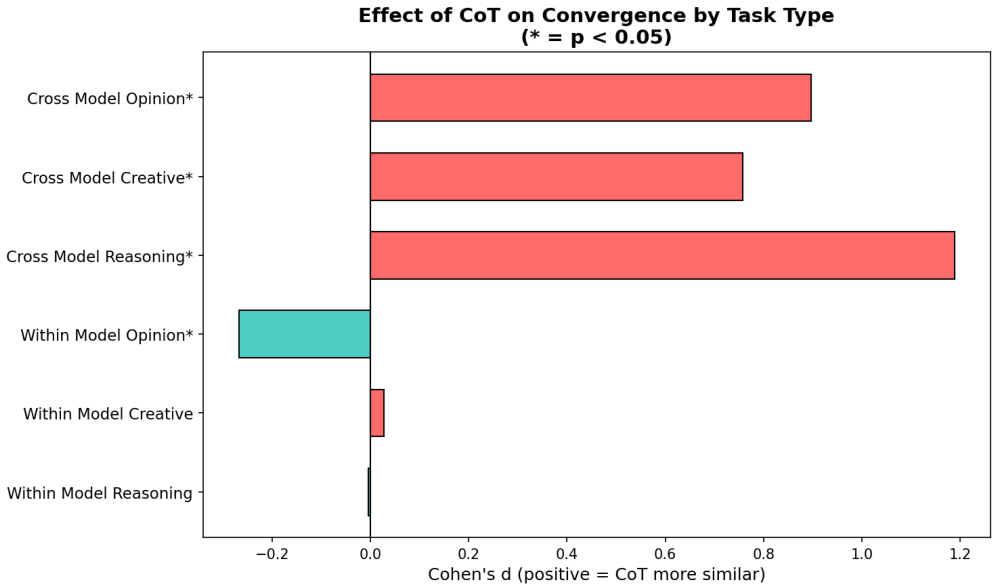

Figure 3: Cohen's d effect sizes for the impact of CoT on convergence. Cross-model effects (positive, right side) are consistently large, while within-model effects (near zero or slightly negative) are negligible. The dashed lines at ± 0.8 mark the conventional threshold for a "large" effect.

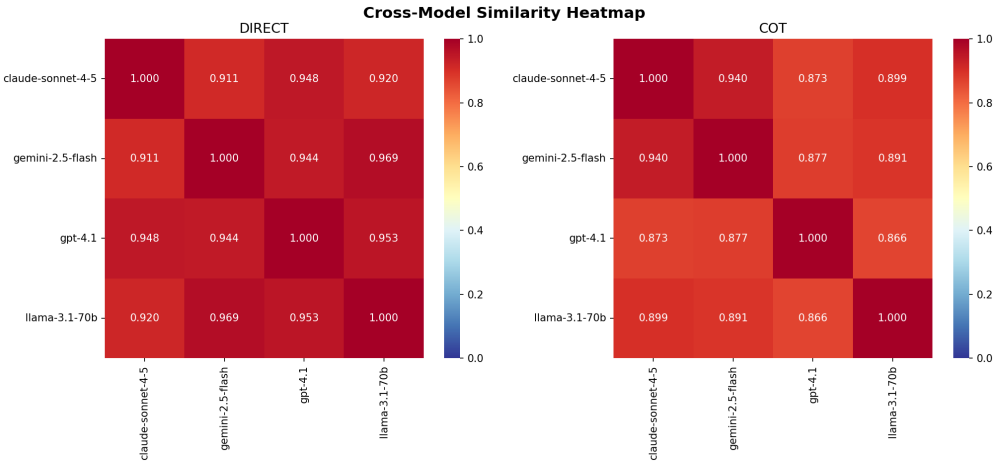

Figure 4: Pairwise similarity heatmap between all model pairs under direct (left) and CoT (right) prompting. Under CoT, all model pairs converge to high similarity, with reduced variance across pairs. The convergence is not driven by any single model pair.

which could inflate embedding similarity. However, this explanation alone is insufficient: if length were the sole driver, *both* within-model and cross-model similarity should increase equally. The fact that only cross-model similarity increases suggests genuine content convergence beyond what length alone would predict.

Convergence in reasoning structure, not content. Manual inspection of responses reveals that CoT outputs across all four models adopt nearly identical structures: numbered steps, bold headers, similar transition phrases ("Let's break this down," "Step 1:"). Direct responses, despite being much shorter, show more structural diversity—some models use bullet points, others give terse answers, others provide brief explanations. CoT appears to impose a shared *format* that sentence embeddings capture as high similarity.

5.2 The Answer Agreement Paradox

Our finding that CoT decreases answer agreement (44.8% $\rightarrow$ 22.9%) despite increasing semantic similarity deserves careful interpretation. We see two potential explanations. First, the longer CoT reasoning chains may introduce more opportunities for errors to compound—small differences in intermediate steps can lead to different conclusions even when the overall reasoning approach is similar. Second, this pattern is consistent with the unfaithful CoT literature [Turpin et al., 2023, Chen et al., 2025]: if models arrive at answers through internal computations that differ from their stated reasoning, the convergence of reasoning traces need not imply convergence of answers.

5.3 Limitations

Sample size. Our study uses 46 questions with 3 samples per condition. While the effect sizes are large and consistent, studies with hundreds of questions would provide greater statistical power and enable finer-grained analysis by task type.

Single CoT variant. We test only zero-shot CoT ("Let's think step by step"). Few-shot CoT, self-consistency decoding, and native reasoning models (e.g., OpenAI o1, DeepSeek R1) may show different convergence patterns. In particular, models trained with reasoning-specific RL may have learned to produce more diverse reasoning paths than prompted CoT.

Four models. Expanding to additional model families (Mistral, Qwen, Cohere, DeepSeek) would strengthen the generalizability of our findings. Our current selection covers the four largest model families but does not include smaller or more specialized models.

Embedding-based metrics. Sentence embeddings capture semantic similarity but may miss functional differences that matter in practice. An LLM-judge approach [Jain et al., 2025] that evaluates whether two responses are *functionally equivalent* (i.e., would lead to the same downstream decisions) could reveal nuances that embedding similarity misses. Additionally, our embedding model (ALL-MINILM-L6-V2) truncates at 256 tokens, which may not capture the full extent of long CoT reasoning chains.

Causal claims. While we observe that CoT increases cross-model convergence, we cannot fully determine the causal mechanism. The convergence could be driven by shared training data, by inherent properties of step-by-step reasoning, or by the specific prompt template we use. Disentangling these factors requires additional experiments, such as varying prompt templates or testing on models with known non-overlapping training data.

5.4 Implications

For AI deployment. Organizations that rely on multiple LLM providers for diversity—for example, generating multiple candidate solutions and selecting among them—should be aware that CoT significantly reduces the effective diversity of their model ensemble. Under direct prompting, models produce meaningfully different responses; under CoT, the responses become nearly interchangeable.

For benchmark design. Evaluations that use CoT may underestimate true inter-model differences. If two models achieve similar scores on a reasoning benchmark with CoT, this may reflect convergent reasoning templates rather than equivalent underlying capabilities. Benchmarks designed to measure model diversity should consider evaluating under both direct and CoT conditions.

For the homogenization debate. Our results provide the first evidence that a specific, modifiable prompting practice—rather than just training data overlap or architectural choices—contributes to LLM output homogenization. This suggests that the convergence phenomenon has multiple interacting causes, and that some degree of homogenization may be reversible by changing how we prompt models.

6 Conclusion

We present the first direct empirical study of Chain of Thought prompting's effect on cross-model output convergence. By prompting four LLMs from different families with and without CoT across 46 questions spanning reasoning, creative, and opinion tasks, we find that CoT acts as a *cross-model homogenizer*: it dramatically increases the similarity of outputs across model families (Cohen's

$d = 0.93$, $p < 0.00001$) while not increasing within-model repetitiveness ($d = -0.074$). The effect is strongest for reasoning tasks ($d = 1.19$) and persists even for creative and opinion tasks ($d = 0.76$ and $d = 0.90$, respectively).

The key takeaway is that CoT makes different models sound alike without making any individual model more predictable—it narrows the space between models while preserving each model's internal variability.

Future work should control for response length by truncating CoT outputs, test additional CoT variants (few-shot, self-consistency, native reasoning models), expand to more model families, and employ functional diversity metrics beyond embedding similarity. Most importantly, understanding *why* CoT homogenizes—whether through shared training data, inherent properties of step-by-step reasoning, or prompt template effects—would inform strategies for preserving model diversity in an era of ubiquitous reasoning.

References

Syed Adnan Afzal et al. Knowing before saying: LLM representations encode chain-of-thought success before completion. *arXiv preprint arXiv:2505.24362*, 2025.

Yanda Chen et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Nishant Jain et al. LLM output homogenization is task dependent. *arXiv preprint arXiv:2509.21267*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 2022.

Benedikt Mirka et al. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

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. In *Advances in Neural Information Processing Systems*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Mika Wenger and Yoed N Kenett. We're different, we're the same: Creative homogeneity across LLMs. *arXiv preprint arXiv:2501.19361*, 2025.

Hanqi Xu and Yining Zhang. Language of thought shapes output diversity in large language models. *arXiv preprint arXiv:2601.11227*, 2026.