Chain-of-Thought Prompting Does Not Uniformly Increase Convergence in Frontier Language Models

Ari Holtzman and Idea-Explorer

Abstract

Chain-of-thought prompting is widely used to raise reasoning accuracy, but it remains unclear whether it also increases convergence in broader model behavior. We ask a focused question: does chain-of-thought make different frontier models converge more in answers, semantic outputs, and persona expression? We run a controlled API-based study with GPT-4.1 and CLAUDE SONNET 4.5 on matched subsets of GSM8K, BBH-DATE, BBH-LOGIC-3OBJ, and PERSONA-CHAT, comparing NO-COT against COT under fixed seeds and standardized parsing. We evaluate exact-match accuracy, cross-model answer agreement, cross-model embedding cosine similarity, persona adherence, and within-model persona stability under repeated stochastic generations. Results are mixed rather than uniformly positive. CoT improves GPT-4.1 performance on GSM8K (0.600 to 0.900) and raises cross-model agreement on GSM8K (0.633 to 0.900), but it does not improve convergence on both BBH subsets. The strongest statistically reliable effect is negative: persona stability for CLAUDE SONNET 4.5 drops from 0.857 to 0.753 ($p = 0.00085$, FDR-adjusted $p = 0.0145$, Cohen's $d = -0.679$). Most other effects are not significant after FDR correction. These findings support a conditional view of CoT: it can improve convergence on some structured reasoning tasks while degrading persona-level consistency, so deployment decisions should optimize for the specific stability axis that matters.

1 Introduction

Reliability in language-model systems depends on more than accuracy. In production settings, teams often care about whether outputs converge across models, prompts, and repeated generations. Chain-of-thought (CoT) prompting is now a standard intervention for reasoning tasks, but its effect on broader convergence remains unclear.

what is missing? Prior studies establish that CoT can improve reasoning performance, especially on multi-step tasks [Wei et al., 2022, Kojima et al., 2022, Wang et al., 2022]. At the same time, faithfulness work shows that verbalized rationales may diverge from true decision factors [Turpin et al., 2023, Ley et al., 2024], and persona work reports instability in self-consistent behavior [Alikhani et al., 2024, Ai et al., 2024, Xu et al., 2024]. Existing evidence therefore does not answer whether CoT improves global convergence across answer, semantic, and persona dimensions under a single matched protocol.

our main question. We examine whether CoT makes frontier models converge more, rather than only score higher on selected benchmarks. We compare NO-COT and COT for GPT-4.1 and CLAUDE SONNET 4.5 across three reasoning datasets and one persona dataset, then evaluate convergence with discrete and embedding-based metrics.

quantitative preview. CoT increases GPT-4.1 accuracy on GSM8K by 30.0 points (0.600 to 0.900) and raises cross-model agreement on GSM8K by 26.7 points (0.633 to 0.900). However, convergence does not improve on BBH subsets, and persona stability for CLAUDE SONNET 4.5 decreases by 10.4 points (0.857 to 0.753), the only effect that remains significant after FDR correction.

Our contributions are:

- We propose TRICONVERGEEVAL, a unified evaluation protocol that measures answer-level, semantic, and persona convergence under matched prompting controls.
- We conduct controlled experiments on real API models and four benchmark families using deterministic subsampling, standardized parsing, and cached execution.
- We show that CoT yields axis-specific effects: improvements in some reasoning settings but a significant degradation in persona stability for CLAUDE SONNET 4.5.
- We provide a practical decision framing: CoT should be selected by target reliability axis, not assumed to be a global convergence enhancer.

Paper organization. Section 2 situates our work in CoT, faithfulness, and persona-consistency literature. Section 3 describes datasets, prompting conditions, and metrics. Section 4 presents empirical results and statistical tests. Section 5 discusses implications and limitations, and Section 6 concludes.

2 Related Work

Chain-of-thought for reasoning performance. CoT prompting substantially improves reasoning accuracy in large models [Wei et al., 2022], including zero-shot variants [Kojima et al., 2022]. Self-consistency extends this idea by sampling multiple reasoning paths and aggregating answers [Wang et al., 2022]. These studies primarily optimize end-task correctness; they do not directly test whether convergence improves across different behavioral axes.

Faithfulness and explanation-behavior mismatch. A key challenge is that generated rationales are often not faithful to internal computation [Turpin et al., 2023]. Follow-up work finds that jointly maximizing faithfulness and accuracy is difficult in practice [Ley et al., 2024]. Faithful CoT methods attempt to reduce this gap [Xiao et al., 2023], but existing evaluations still focus on explanation quality or answer quality rather than global convergence.

Evaluation infrastructure and variance across tasks. Broad benchmark efforts such as Chain-of-Thought Hub show strong task- and model-dependent variation in CoT outcomes [Fu et al., 2023]. This motivates controlled, cross-axis measurement rather than single-metric claims.

Persona consistency and self-knowledge. Persona-steered generation studies show that conditioning can alter behavior and amplify unwanted shifts [Alikhani et al., 2024]. Related work on self-knowledge and self-cognition reports incomplete alignment between stated traits and generated actions [Ai et al., 2024, Xu et al., 2024]. Our work connects this line to CoT by testing whether explicit reasoning prompts stabilize or destabilize persona expression.

Positioning. Unlike prior work that isolates one outcome (e.g., accuracy, faithfulness, or persona drift), we evaluate CoT under one protocol across answer agreement, semantic similarity, and persona stability. This unified perspective is necessary to decide whether CoT improves reliability broadly or only in narrow task settings.

3 Methodology

what do we test? We test whether CoT increases convergence across three dimensions: (1) reasoning answers, (2) semantic output representations, and (3) persona behavior. We compare NO-COT and COT with identical datasets and fixed controls.

3.1 Experimental Setup

We evaluate two API models: GPT-4.1 (OpenAI) and CLAUDE SONNET 4.5 (via OpenRouter's OpenAI-compatible interface). We use GSM8K (1,319 items), BBH-DATE (250), BBH-LOGIC-3OBJ (250), and PERSONA-CHAT (1,000 validation dialogues). For compute and cost control, we apply deterministic subsampling (seed 42) to 30 examples per dataset while keeping item identities fixed across all conditions.

No model training or fine-tuning is performed. We run evaluation only. Reasoning prompts use temperature 0.0. Persona prompts use temperature 0.7 with two repeated generations per item to estimate stability under stochastic decoding. Maximum generation length is 350 tokens.

3.2 Prompt Conditions and Parsing

The NO-COT condition requests direct final outputs. The COT condition requests explicit step-by-step reasoning before a structured final field. For reasoning tasks, outputs are parsed from `FINAL_ANSWER:`; for persona tasks, outputs are parsed from `RESPONSE:`. Gold labels are normalized using task-specific rules: numeric extraction after `####` for GSM8K and option-letter extraction for BBH tasks.

3.3 Convergence Metrics

We report:

- **Accuracy**: exact normalized match to gold labels for GSM8K and BBH tasks.
- **Cross-model answer agreement**: exact-match agreement between GPT-4.1 and CLAUDE SONNET 4.5 on each item.
- **Cross-model semantic convergence**: cosine similarity between model output embeddings.
- **Persona adherence**: cosine similarity between response embeddings and persona-profile embeddings.
- **Persona stability**: within-model cosine similarity across repeated responses for the same persona prompt.

Embeddings are computed with `text-embedding-3-small` and cached for reproducibility.

3.4 Statistical Analysis

For each paired comparison (COT minus NO-COT), we test normality of paired differences with Shapiro–Wilk. When normality holds, we use paired t-tests; otherwise, Wilcoxon signed-rank tests. We report raw p values, 95% confidence intervals of mean paired differences, and Cohen's d for paired effects. Because we test multiple hypotheses, we apply Benjamini–Hochberg FDR correction with $\alpha = 0.05$.

3.5 Baselines and Reproducibility

Our primary baseline is NO-COT. CoT is the treatment condition. Infrastructure includes request-hash caching, retry/backoff, deterministic sampling manifests, and fixed seeds. The run produced 600 API outputs in approximately 23 minutes; analysis required approximately 7 seconds.

4 Results

Main outcomes. Table 1 shows that CoT effects are task dependent. For GPT-4.1, CoT improves GSM8K accuracy from 0.600 to 0.900 and BBH-LOGIC-3OBJ from 0.933 to 1.000, but lowers BBH-DATE from 0.333 to 0.267. For CLAUDE SONNET 4.5, GSM8K remains unchanged at 0.900, while BBH subsets decline (0.900 to 0.833 on BBH-DATE; 1.000 to 0.933 on BBH-LOGIC-3OBJ).

Convergence beyond accuracy. Table 2 indicates that cross-model agreement increases only on GSM8K (0.633 to 0.900) and is unchanged on both BBH tasks. Persona adherence remains nearly flat for GPT-4.1 (0.385 to 0.389) and decreases for CLAUDE SONNET 4.5 (0.439 to 0.416). Persona stability decreases for both models, with a pronounced drop for CLAUDE SONNET 4.5 (0.857 to 0.753).

Statistical tests. Most effects do not survive FDR correction. Table 3 summarizes the strongest signals: GPT-4.1 GSM8K gain and GSM8K agreement gain are near-significant after correction, while the CLAUDE SONNET 4.5 persona-stability drop is significant ($p = 0.00085$, FDR $p = 0.0145$, $d = -0.679$).

Task	Model	NO-COT	COT
GSM8K	GPT-4.1	0.600	**0.900**
GSM8K	CLAUDE SONNET 4.5	**0.900**	**0.900**
BBH-DATE	GPT-4.1	**0.333**	0.267
BBH-DATE	CLAUDE SONNET 4.5	**0.900**	0.833
BBH-LOGIC-3OBJ	GPT-4.1	0.933	**1.000**
BBH-LOGIC-3OBJ	CLAUDE SONNET 4.5	**1.000**	0.933

Table 1: Reasoning accuracy by task, model, and prompting condition. Best value per row is in **bold**. CoT helps GPT-4.1 on GSM8K and BBH-LOGIC-3OBJ, but not BBH-DATE; effects are mixed for CLAUDE SONNET 4.5.

Metric	Slice	NO-COT	COT
Cross-model answer agreement	GSM8K	0.633	**0.900**
Cross-model answer agreement	BBH-DATE	**0.300**	**0.300**
Cross-model answer agreement	BBH-LOGIC-3OBJ	**0.933**	**0.933**
Persona adherence	GPT-4.1	0.385	**0.389**
Persona adherence	CLAUDE SONNET 4.5	**0.439**	0.416
Persona stability	GPT-4.1	**0.758**	0.735
Persona stability	CLAUDE SONNET 4.5	**0.857**	0.753

Table 2: Agreement and persona metrics across prompting conditions. CoT improves agreement only on GSM8K and reduces persona stability for both models, with a large drop for CLAUDE SONNET 4.5.

Error patterns. On BBH-DATE, failures are dominated by temporal interpretation and option-extraction mistakes. On GSM8K, GPT-4.1 NO-COT errors often show plausible intermediate work but incorrect final arithmetic values. In persona generation, CoT outputs more meta-explanatory text, which likely reduces stylistic consistency and contributes to lower stability.

5 Discussion

Interpretation. Our findings reject a simple claim that CoT increases convergence globally. Instead, CoT behaves as a targeted intervention: it can align models on specific structured tasks (GSM8K), while leaving other tasks unchanged or worse and reducing persona stability. This pattern is consistent with prior evidence that rationale quality and decision faithfulness can diverge [Turpin et al., 2023, Ley et al., 2024].

Why the trade-off may occur. CoT imposes a reasoning format that can regularize final answers for arithmetic tasks, which may explain stronger agreement on GSM8K. In persona-conditioned dialogue, the same format introduces explanatory content that competes with role expression, reducing response consistency across repeats.

Practical implications. Deployment should optimize the reliability target explicitly. If the goal is higher arithmetic agreement, CoT may help. If the goal is stable persona behavior in conversational agents, NO-COT or alternative controls may be safer defaults. Multi-model systems should track axis-specific reliability rather than reporting a single aggregate "consistency" score.

Limitations. First, subset size is small (30 per dataset slice), limiting power for moderate effects. Second, persona metrics use embedding proxies rather than human judgments. Third, only two

Effect (CoT−NoCoT)	Raw p	FDR p	Cohen's d	95% CI
GPT-4.1 GSM8K accuracy	0.0067	0.0566	0.562	[-0.001, 0.601]
GSM8K agreement	0.0114	0.0647	0.485	[0.032, 0.501]
CLAUDE SONNET 4.5 persona stability	**0.00085**	**0.0145**	-0.679	[-0.159, -0.049]

Table 3: Representative hypothesis-test outcomes from the full test suite. The persona-stability drop for CLAUDE SONNET 4.5 is the only robust effect after FDR correction.

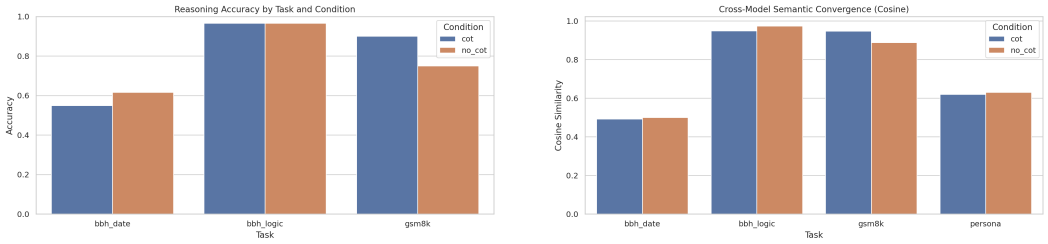

Figure 1: Left: accuracy by task and prompting condition. Right: embedding-based semantic convergence across tasks. CoT gains are concentrated in GSM8K and do not generalize uniformly.

models are tested, and both are closed APIs with no access to internal activations. Fourth, two repeats for persona stability provide only a minimal variance estimate.

Broader implications and risk. CoT can improve perceived rigor of outputs while masking instability in behavior dimensions users care about. Evaluations for safety-critical or user-facing applications should therefore include explicit persona and semantic consistency tests, not only correctness on benchmark labels.

6 Conclusion

We presented a unified evaluation of whether chain-of-thought prompting increases convergence in frontier language models across answer, semantic, and persona dimensions. Across GPT-4.1 and CLAUDE SONNET 4.5, CoT effects were mixed: beneficial on selected reasoning outcomes, neutral on others, and harmful for persona stability in the strongest statistically supported result.

The key takeaway is straightforward: CoT is not a universal convergence mechanism. It is a conditional tool whose value depends on the specific reliability axis and task distribution.

Future work should increase sample sizes (e.g., at least 100 items per slice), include self-consistency decoding to disentangle reasoning-format and decoding effects, and add human-rated persona consistency with inter-annotator agreement. Extending evaluation to additional models and long-horizon multi-turn settings will further clarify when CoT improves reliability versus when it introduces instability.

References

Yiming Ai et al. Is self-knowledge and action consistent or not: Investigating the coherence of large language models in theory and practice. *arXiv preprint arXiv:2402.14679*, 2024.

Malihe Alikhani et al. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*, 2024.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

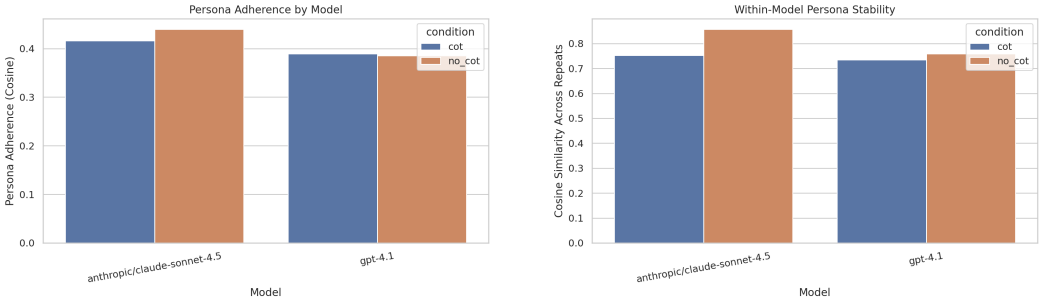

Figure 2: Persona outcomes under NO-COT and COT. Adherence shifts are modest, but stability declines under CoT, especially for CLAUDE SONNET 4.5.

Dan Ley, Oshin Agarwal, Kalpesh Krishna, and Eric Wallace. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 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 preprint arXiv:2305.04388*, 2023.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Chi, Fei Xia, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Yaozong Xiao, Denny Zhou, Yuxian Gu, Yiming Yang, Weinan Zhang, and Jiawei Han. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

Weichen Xu et al. Self-cognition in large language models: An exploratory study. *arXiv preprint arXiv:2407.01505*, 2024.