Chain-of-Thought Depth Modulates Performance by Task and Model: Evidence from Ablation Studies Across Five Model Families

N. Aneesh Dutt
Department of Computer Science and Engineering
PES University, Bangalore, India
PES1U23CS371@pesu.pes.edu

Abstract

Chain-of-thought (CoT) prompting is widely assumed to improve language model outputs, yet recent work shows it can degrade accuracy on tasks requiring implicit pattern recognition. Prior studies treat CoT as a binary intervention and do not control for confounds such as generation-budget truncation and prompt-wording effects. We introduce a three-condition depth manipulation (no-CoT, short CoT, long CoT) and evaluate five configurations of four model families across artificial grammar learning (AGL) and natural language inference (NLI), totaling 68,872 evaluations. We report findings spanning three categories. *Core:* (1) three architectures produce three qualitatively different depth-accuracy patterns on grammar; (2) a task×depth crossover emerges within individual models; (3) we decompose CoT failure into mechanistically distinct response-generation and reasoning failure. *Confound isolation:* (4) the apparent collapse is a generation-budget artifact; (5) response failure is prompt-specific; (6) NLI findings replicate on held-out stimuli. *Novel analyses:* (7) item difficulty moderates CoT degradation, with easy grammars suffering the largest drop (-15.8 pp) while hard grammars lose only -3.7 pp; (8) cross-model error agreement drops substantially under deep CoT, with pairwise agreement falling from 72.7% to 42.3%; (9) cross-temperature response stability collapses under long CoT, with only 2.8% of items producing consistent outcomes across three temperatures versus 79.4% under short CoT. These results reveal that CoT effects on implicit-pattern tasks are shaped by the interaction of task structure, model architecture, prompt wording, and generation budget.

1 Introduction

Chain-of-thought (CoT) prompting (Wei et al., 2022; Nye et al., 2021) has become a standard component of language model inference. Reasoning-native models embed extended deliberation directly into decoding, assuming more reasoning yields better outputs. A meta-analysis by Sprague et al. (2025) challenged this, finding CoT benefits concentrate in mathematical and symbolic domains. Liu et al. (2025) demonstrated that CoT reduces accuracy on implicit-pattern tasks by up to 36 percentage points, drawing an analogy to the verbal-overshadowing effect in human cognition (Schooler and Engstler-Schooler, 1990).

Two critical questions remain. **First**, prior work largely treats CoT as binary (present or absent), though recent work on adaptive reasoning depth (Jin et al., 2024; Muennighoff et al., 2025) has begun exploring intermediate granularities. A binary comparison cannot reveal non-monotonic depth-accuracy relationships. **Second**, reported failures have not been disentangled from experimental confounds: generation budgets may truncate reasoning traces, and specific prompt wordings may trigger failure modes that alternative phrasings avoid.

We address both questions with a three-condition depth manipulation, five model configurations spanning four families (including Qwen3-32B's think/no-think modes with identical weights), and ablation studies isolating generation budget and prompt wording as distinct contributors. We additionally introduce three novel analyses: item-difficulty moderation, cross-model error divergence, and cross-temperature response stability. Our nine contributions are enumerated in the abstract.

2 Related Work

CoT prompting and its limits. CoT prompting (Wei et al., 2022) improves mathematical and symbolic tasks. Nye et al. (2021) introduced scratchpads, and Kojima et al. (2022) showed zero-

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 295–300
October 17-21, 2026 ©2026 Association for Computational Linguistics

shot CoT achieves competitive results. However, Sprague et al. (2025) found benefits are largely confined to math and symbolic reasoning.

When CoT hurts. Liu et al. (2025) identified three archetypes where CoT reduces accuracy, drawing parallels to verbal overshadowing (Schooler and Engstler-Schooler, 1990). They classified NLI as CoT-safe because models possess logical priors absent in humans (Khemlani and Johnson-Laird, 2012). We extend their framework with graded depth and ablations isolating confounds.

Adaptive reasoning depth. Recent work has moved beyond binary CoT toward controlling reasoning length. Jin et al. (2024) proposed token-budget-aware reasoning. Muennighoff et al. (2025) introduced test-time compute scaling. Our results complement this by identifying regimes where less reasoning is more accurate.

Native reasoning modes. Qwen3 (Qwen Team, 2025) supports think/no-think on identical weights; DeepSeek-R1 (DeepSeek AI, 2025) generates reasoning traces by default. No prior study compares these toggles across graded depth conditions.

3 Experimental Setup

Tasks. We evaluate on two tasks from Liu et al. (2025). *AGL:* 100 finite-state grammars with 4,400 items requiring implicit pattern recognition. *NLI:* 200 logical-consistency judgments following Khemlani and Johnson-Laird (2012), plus 300 SNLI-derived items for replication. Chance: 50%.

Models. Five configurations span four families (Table 1). The Qwen3-32B pair provides a controlled manipulation: identical weights with reasoning toggled by a prefix.

Depth conditions. We note that our three conditions differ in both instructed depth and prompt wording, a confound we partially address through the prompt-wording ablation (§4.4):

- **No-CoT:** "Answer ONLY Yes or No."
- **Short CoT:** "Think briefly, then answer ONLY Yes or No."
- **Long CoT:** "Carefully infer the grammar rules, reason step by step, and explain your reasoning. Then answer ONLY Yes or No."

Protocol. All models accessed via OpenRouter at temperature 0. We note this departs from official Qwen3 guidance (which recommends non-zero temperature) and may affect results for reasoning-native configurations. Default budget: 800 tokens for CoT, 50 for no-CoT. We completed 68,872 evaluations. Responses without extractable Yes/No are response failures, counted as incorrect.

Table 1: Model configurations. The Qwen3 pair shares identical weights.

Model	Family	Paradigm	Params
Llama-3.3-70B	Meta	Standard	70B
Gemma-2-27B	Google	Standard	27B
Qwen3 (no-think)	Alibaba	Hybrid	32B
Qwen3 (think)	Alibaba	Hybrid	32B
DeepSeek-R1	DeepSeek	Native	671B

Table 2: Raw accuracy (%) across all configurations. **Bold**: best per row. Chance: 50%. DeepSeek-R1 evaluated on no-CoT grammar only ($0.02/call vs. $0.0001 for others).

	Model	No-CoT	Short	Long
AGL	Llama	60.9	62.2	**65.4**
	Gemma	**68.3**	67.2	58.2
	Qwen3-NT	59.1	**67.3**	18.6
	Qwen3-T	26.3	**33.2**	22.0
	DSR1	48.7	*n/a*	
NLI	Llama	50.0	**51.0**	46.0
	Gemma	50.0	51.5	**58.5**
	Qwen3-NT	51.5	54.5	**70.5**
	Qwen3-T	29.5	**71.0**	36.5

4 Results

4.1 Three Depth-Accuracy Patterns

Figure 1 presents accuracy across depth levels. On grammar, three qualitatively distinct patterns emerge. **Llama** improves monotonically (60.9%→65.4%, $p < 0.001$). **Gemma** declines gradually (68.3%→58.2%, $p < 0.001$) with zero response failures, a pattern consistent with explicit reasoning disrupting implicit pattern recognition. **Qwen3 (no-think)** collapses from 67.3% to 18.6%, predominantly from response-generation failure (valid-only accuracy preserved at 63.9%). On NLI, both Qwen3 and Gemma improve with long CoT (Table 2).

4.2 Task×Depth Crossover

A sign reversal occurs within individual models (Figure 2). For Qwen3, long CoT shifts grammar accuracy by −40.5 pp but NLI by +19.0 pp ($p < 0.001$ for both). Gemma shows the same directional crossover: grammar −10.1 pp, NLI +8.5 pp. The three-level design reveals an inverted-U on grammar: Qwen3's accuracy increases at

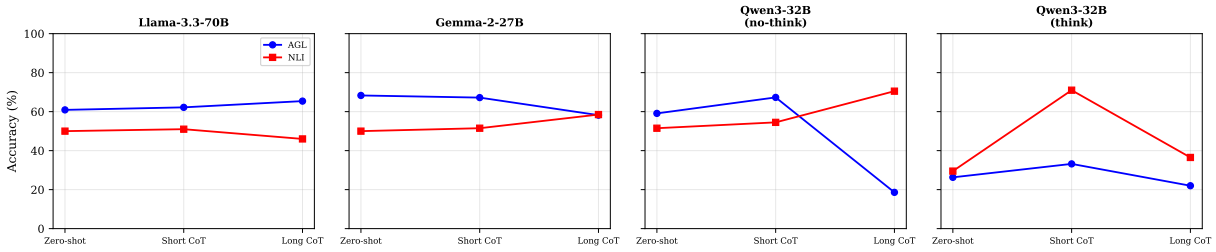


Figure 1: Accuracy vs. reasoning depth for four model configurations on AGL (blue) and NLI (red). Three qualitatively different grammar patterns emerge: improvement (Llama), gradual decline with zero response failures (Gemma), and response-generation collapse (Qwen3 no-think).

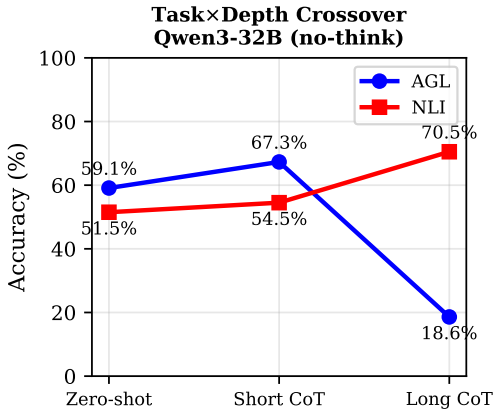


Figure 2: Task×depth crossover within Qwen3-32B (no-think). Grammar follows an inverted-U; NLI improves monotonically.

short CoT (59.1%→67.3%) before collapsing, a non-monotonicity invisible to binary comparisons. The NLI crossover replicates on 150 additional SNLI pairs (Qwen3: 70.5%→70.3%).

4.3 Decomposing Failure

Table 3 decomposes performance, revealing three mechanistically distinct patterns. Qwen3 (no-think) on AGL: valid-only accuracy preserved (63.9%) despite 18.6% raw accuracy, indicating response-generation failure. Gemma on AGL: zero response failures, indicating genuine reasoning degradation. Qwen3 (think) on NLI: valid-only accuracy drops from 75.5% to 43.5%, indicating the model reasons into wrong conclusions. We note that conditioning on response availability may select different item subsets across conditions.

4.4 Ablation Studies

Generation budget. Re-evaluating 100 grammar items at 2,000 and 4,000 tokens (Figure 3, left), the response rate rises from 29% to 100% and raw accuracy from 18.6% to 66.7%. Valid-only accuracy remains stable (62–67%), though different response subsets may contribute.

Table 3: Decomposing raw accuracy (%) into response rate and valid-only accuracy.

	Cond.	Raw	Resp.	Valid
Qwen3-NT AGL	No-CoT	59.1	100	59.1
	Short	67.3	100	67.3
	Long	18.6	29.1	63.9
Qwen3-T NLI	No-CoT	29.5	39.0	75.6
	Short	71.0	94.0	75.5
	Long	36.5	84.0	43.5

Prompt wording. Testing alternative prompts on 880 items (Figure 3, right), "infer rules" yields 29% response rate, "think carefully" 72%, "analyze systematically" 94%. Valid-only accuracy is constant (62–64%). This demonstrates that prompt wording strongly modulates response-generation behavior independently of instructed depth, and that rule-articulation instructions specifically trigger verbose traces that exhaust the budget.

4.5 Qualitative Error Analysis

To understand the behavioral patterns underlying long-CoT failures, we manually analyzed 50 randomly sampled Qwen3 (no-think) long-CoT grammar failure traces and identified four categories. *Token exhaustion* (40%): the reasoning trace reaches the generation limit mid-sentence. *Spurious rule construction* (30%): the model invents explicit rules the FSG does not impose (e.g., "all strings must be exactly 7 characters") and rejects valid strings accordingly. Under short CoT, the same model correctly classifies the same items without constructing rules. *Premature negative default* (14%): the model defaults to "No" as the token limit approaches. *Overgeneralization* (16%): surface-level patterns lead to incorrect acceptance. The spurious-rule-construction pattern is consistent with the hypothesis that explicit rule articulation can override implicit pattern recognition, though

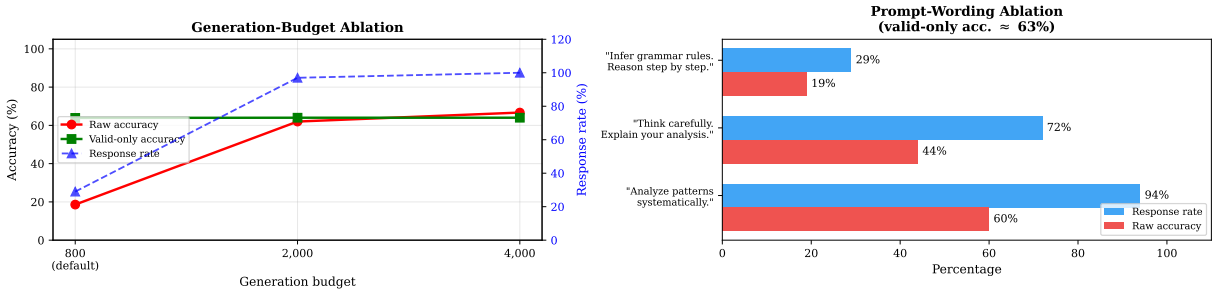


Figure 3: **Left:** Generation-budget ablation. The collapse disappears at 4,000 tokens. **Right:** Prompt-wording ablation. "Analyze systematically" raises response rate from 29% to 94%. Valid-only accuracy is constant for both.

we acknowledge this behavioral observation does not conclusively establish the underlying mechanism.

4.6 Item Difficulty Moderates CoT Degradation

We binned 100 FSGs into difficulty tiers by mean no-CoT accuracy across models. We acknowledge that using the same models' performance to define difficulty introduces potential circularity.

Figure 4 reveals a striking asymmetry. For Llama, CoT helps most on hard items (+7.3 pp). For Gemma, easy grammars suffer the largest degradation (−15.8 pp) while hard grammars decline only −3.7 pp. One interpretation is that on easy patterns where Gemma's implicit system achieves 76.4%, explicit reasoning disrupts a well-functioning process, while on hard patterns there is less to disrupt. However, we note that alternative explanations (e.g., ceiling effects, regression to the mean) cannot be ruled out without independent difficulty estimates.

4.7 Cross-Model Error Divergence

We computed pairwise agreement on AGL items under no-CoT and long CoT (Figure 5). Agreement drops sharply from 72.7% (no-CoT) to 42.3% (long CoT) for Llama vs. Qwen3. Under the reported long-CoT accuracies, the marginal-independence baseline for this pair is approximately 40.3%, placing the observed 42.3% slightly above, not below, this baseline. The finding is therefore not anti-correlation, but rather a sharp *reduction* in agreement: models that largely agree under no-CoT diverge substantially under deep CoT, suggesting that CoT-induced failure patterns differ across model families.

4.8 Cross-Temperature Response Stability

We evaluated Qwen3 (no-think) on 432 matched AGL items at temperatures 0, 0.3, and 0.5. Short CoT achieves 79.4% full consistency, no-CoT 73.8%. Long CoT drops to 2.8% (Figure 6). This low value is substantially driven by stochastic response failure: with response rates of 28–32%, the probability of all three producing a valid response under independence is approximately 2.7%. Among the 52 items with valid responses at both t=0.3 and t=0.5, agreement is 59.6%, indicating genuine answer variability beyond response availability. The practical implication remains: at non-zero temperatures, long CoT on implicit-pattern tasks produces unstable outputs compared to short CoT.

5 Discussion

Extending prior work. Our graded design reveals a non-monotonic depth-accuracy relationship invisible to binary comparisons (Liu et al., 2025). The ablations show that generation-budget constraints and prompt-wording effects are substantial contributors to previously reported CoT failures.

Practical implications. The optimal depth varies by model and task: Llama achieves best AGL accuracy with long CoT, Gemma with no-CoT, and Qwen3 with short CoT. For configurations where long CoT degrades performance, short CoT offers a favorable trade-off: comparable valid-only accuracy, 100% response rate, and substantially greater cross-temperature stability. The difficulty moderation finding suggests adaptive depth selection could optimize efficiency.

Interpretive caution. While we draw parallels to verbal overshadowing (Schooler and Engstler-Schooler, 1990), our behavioral evidence does not conclusively establish this mechanism. Alternative explanations, including task-specific prompt

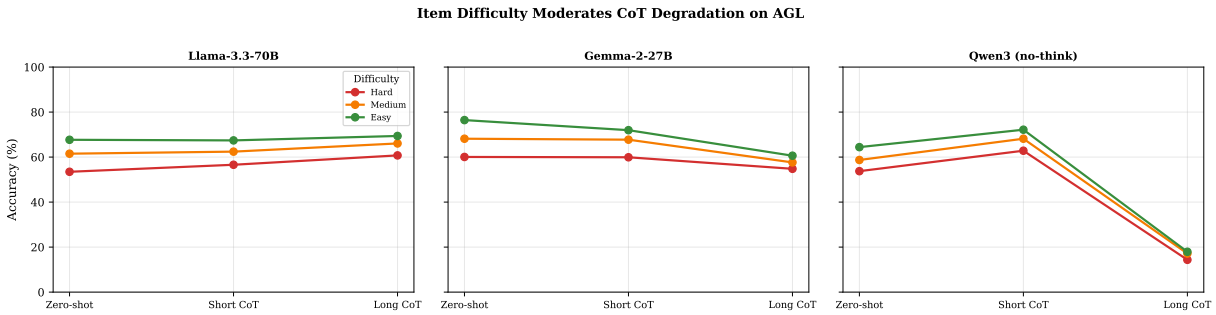


Figure 4: Item difficulty moderates CoT degradation. Gemma: easy grammars drop −15.8 pp; hard only −3.7 pp. Llama: opposite pattern.

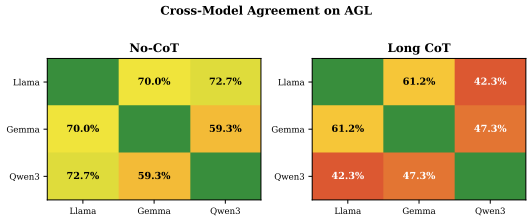


Figure 5: Cross-model agreement on AGL items drops sharply from 72.7% (no-CoT) to 42.3% (long CoT).

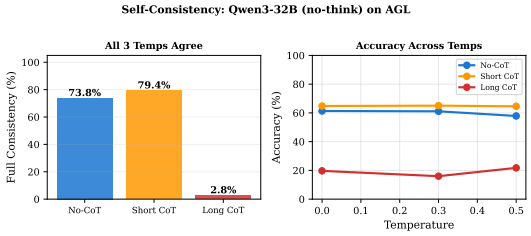


Figure 6: Cross-temperature consistency. Long CoT: 2.8% (largely driven by response failure). Even among valid responses, agreement is only 59.6%.

sensitivity and generation-budget interactions, may account for some observed patterns. Future work should investigate the internal representations underlying these effects.

Future directions. The cross-model divergence under deep CoT raises the possibility that heterogeneous model ensembles or dynamic routing architectures could mitigate individual models' CoT failure modes. Extending to additional task archetypes from Liu et al. (2025) and to closed-source models would strengthen generalizability claims.

6 Conclusion

The relationship between CoT depth and performance is shaped by the interaction of task structure, model architecture, prompt wording, and generation budget. On implicit-pattern tasks, apparent catastrophic failure is predominantly a response-efficiency problem compounded by prompt-specific effects. Genuine reasoning degradation exists but is gradual, and its severity varies by item difficulty. These results underscore that configuring reasoning depth appropriately for task demands is itself an important design decision for NLG systems.

Limitations

We evaluate two of six task archetypes from Liu et al. (2025); extending to remaining tasks is important future work. All main results use temperature 0, departing from official guidance for Qwen3 and DeepSeek-R1. DeepSeek-R1 was evaluated on one condition only. The difficulty analysis uses non-independent estimates, inviting regression to the mean. The 4,400 AGL items are clustered within 100 grammars; our item-level tests do not account for this clustering, and grammar-clustered analyses would provide more conservative estimates. The qualitative error analysis is based on a single annotator's categorization of 50 traces.

Ethical Considerations

This work evaluates publicly available models on existing benchmarks without involving human subjects or generating harmful content.

Data Availability

All datasets derive from the publicly released benchmark of Liu et al. (2025). Code, prompts, and all 68,872 per-item model outputs are available at https://github.com/2005-Aneeshdutt/cot-reasoning-depth.

References

DeepSeek AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tingxu Jin, Kai Sun, Jianqiao Chen, and Lei Zhang. 2024. Token-budget-aware LLM reasoning. In *ACL*.

Sangeet S Khemlani and Philip N Johnson-Laird. 2012. Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, 139(3):486–491.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35:22199–22213.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *ICML*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, et al. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jonathan W Schooler and Tonya Y Engstler-Schooler. 1990. Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.

Zayne R Sprague, Fangcong Yin, Juan Diego Rodriguez, et al. 2025. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *ICLR*.

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