When Thinking Costs More Than It Helps: Single-Pass Competence Predicts Chain-of-Thought Energy Waste

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) prompting has become a default strategy in deployed language model systems, yet its energy cost relative to quality benefit has not been measured systematically across model families on real hardware. We report 8,421 NVML-instrumented inference runs on A100 GPUs, covering six models from five families (Qwen2.5-7B/14B, Llama-3.1-8B, Mistral-7B, Phi-3-Mini-3.8B, Gemma-2-9B) on 86 tasks spanning retrieval, reasoning, and tool-use. A model's *single-pass quality* (how well it answers without chain-of-thought) is the dominant predictor of whether CoT helps or hurts. The relationship is linear across all 18 model–family combinations ($r = -0.73$, $p < 0.001$). Three of six models get *worse* with CoT overall, including the smallest (Phi-3-Mini, $\Delta q = -10.9$ pp) and the largest (Qwen-14B, $\Delta q = -1.6$ pp), not because of size, but because both already answer well without thinking. CoT degrades tool-use accuracy for every model tested (-13 to -53 pp). Where CoT helps most, on reasoning tasks for weaker models, the energy cost runs 2–61$\times$ higher than single-pass, and a token budget sweep (5,160 runs) shows 34–40% of that energy is spent past the quality plateau. Failure analysis of 175 CoT regressions reveals tool-use format disruption behind 54% of cases. These findings motivate headroom-aware CoT routing: invoke reasoning only when single-pass accuracy falls below $\sim$0.83.

1 Introduction

Chain-of-thought prompting (Wei et al., 2022) and its descendants (self-consistency (Wang et al., 2023), self-refinement (Madaan and Collaborators, 2023), multi-agent debate (Du et al., 2024)) have become load-bearing components of production AI pipelines (Zaharia and Collaborators, 2024). The bet is simple: more reasoning tokens, better answers. Test-time compute scaling results (Snell et al., 2025; OpenAI, 2024; DeepSeek AI, 2025) have reinforced this, at least on math benchmarks.

We wanted to check whether this bet pays off in general, and what it costs in energy. So we ran six models from five different families through 86 tasks with GPU-level energy metering, comparing direct-answer (single-pass) responses against chain-of-thought responses.

The result is a clear pattern, but not the one we expected. Model size does not predict CoT utility. *Single-pass quality does.* Phi-3-Mini, our smallest model at 3.8B parameters, suffers the worst overall CoT regression (-10.9 pp) because it already answers most questions correctly without reasoning. Llama-3.1-8B, a mid-sized model, benefits the most ($+9.3$ pp) because its single-pass reasoning accuracy is lowest. Qwen-14B, our largest model, regresses under CoT (-1.6 pp) for the same reason as Phi-3: high single-pass accuracy, not large size. Across all 18 model–family cells, the linear relationship $\Delta q = -1.29 \times q_{\text{SP}} + 1.07$ holds with $r = -0.73$.

Contributions. (i) Evidence from six models and five families that single-pass quality headroom ($r = -0.73$, $N = 18$) predicts CoT utility, where model size and family do not. (ii) Energy-quality profiles of CoT on A100 hardware, showing 2–61$\times$ energy multipliers and quality regressions in 11 of 18 cells. (iii) Token budget saturation analysis revealing 34–40% wasted CoT energy beyond the quality plateau. (iv) A failure taxonomy of 175 regressions: 54% from tool-use format disruption, 18% from retrieval second-guessing.

2 Related Work

When does CoT help? Sprague et al. (2025) found that CoT primarily benefits math and symbolic reasoning, with diminishing or negative returns elsewhere. Our results put a number on this:

CoT gain crosses zero at $q_{\text{SP}} \approx 0.83$. Damani and Collaborators (2025) proposed input-adaptive reasoning allocation but did not provide concrete thresholds; our token budget analysis supplies them.

Verbosity bias. Post-training biases models toward longer outputs (Singhal and Collaborators, 2024). Our data adds a wrinkle: the 14B model generates roughly the same number of reasoning tokens as the 7B (∼196 vs. ∼193) but extracts less quality per token. The issue appears to be what gets said, not how much.

Test-time compute scaling. Snell et al. (2025) showed that test-time compute can outperform parameter scaling under certain conditions. Our measurements bound where this holds in practice: it works when single-pass quality is low (Llama-8B reasoning: +30 pp) and backfires when single-pass quality is high (Phi-3 reasoning: −5.0 pp). Zhou and Collaborators (2026) proposed paradigm routing; our headroom metric offers a concrete selection criterion.

Energy measurement. We follow the NVML methodology recommended by Fernandez et al. (2025) and adopt the energy waste framing of Anonymous (2026). Energy multipliers are computed as $\bar{E}_{\text{CoT}}/\bar{E}_{\text{SP}}$ (ratio of family-level means). Per-task ratios tend to be larger due to Jensen's inequality; we report the conservative number throughout.

3 Experimental Setup

3.1 Models and Hardware

All experiments run on NVIDIA A100-80GB SXM GPUs. We measure energy using `nvmlDeviceGetTotalEnergyConsumption` cumulative counters rather than polled power readings (which miss short bursts), following the protocol of Yang and Collaborators (2023). Each experiment begins with a 5-step warm-up phase, and we subtract idle power draw (50 W baseline) from all measurements. Models are served through vLLM with greedy decoding (temperature 0). We test six models spanning five families: Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Qwen), Llama-3.1-8B-Instruct (Meta), Mistral-7B-Instruct-v0.3 (Mistral AI), Phi-3-Mini-4K-Instruct at 3.8B (Microsoft), and Gemma-2-9B-IT (Google).

Table 1: SP vs. CoT quality and energy across six models and three task families. E_{mult}: energy multiplier. Bold: CoT makes things worse.

Model	Family	SP q	CoT q	Δq	E_{m}
Qwen-7B	Retr.	.903	.914	+.011	40×
	Reas.	.800	.917	+.117	52×
	Tool	1.00	**.733**	**−.267**	15×
Qwen-14B	Retr.	.935	.957	+.022	34×
	Reas.	.875	.925	+.050	61×
	Tool	.978	**.711**	**−.267**	15×
Llama-8B	Retr.	.968	**.903**	**−.065**	25×
	Reas.	.675	.975	+.300	27×
	Tool	.867	**.733**	**−.133**	11×
Mistral-7B	Retr.	.839	**.774**	**−.065**	3×
	Reas.	.600	.658	+.058	3×
	Tool	1.00	**.467**	**−.533**	3×
Phi-3	Retr.	.968	**.935**	**−.032**	5×
	Reas.	.975	**.925**	**−.050**	2×
	Tool	1.00	**.578**	**−.422**	7×
Gemma-9B	Retr.	.968	**.925**	**−.043**	15×
	Reas.	.700	.900	+.200	25×
	Tool	1.00	**.600**	**−.400**	10×

3.2 Tasks and Modes

Our benchmark covers 86 tasks: 31 retrieval (factual lookup), 40 reasoning (arithmetic, logic, probability), and 15 tool-use (code generation, SQL queries, JSON formatting). Quality is binary exact-match, averaged over 3 replications. We compare two prompting modes: **single-pass (SP)**, which instructs the model to answer directly with no explanation, and **chain-of-thought (CoT)**, which asks the model to think step by step before giving a final answer marked with `ANSWER:`.

3.3 Experiments

We ran three experimental designs totaling 8,421 measured inferences. (1) **SP vs. CoT comparison**: 86 tasks × 2 modes × 3 reps = 516 runs per model across all six models (3,096 runs). (2) **Token budget sweep**: 86 tasks × 10 budgets (32–1,024 tokens) × 3 reps = 2,580 runs per model on both Qwen models (5,160 runs). (3) **Failure classification**: manual analysis of every case where SP succeeded and CoT failed, across all models.

4 Results

4.1 Single-Pass Quality Predicts CoT Gain

Table 1 breaks down quality and energy for all 18 model–family cells. Quality regressions (bold) show up in 11 of them. Figure 1 plots the relationship.

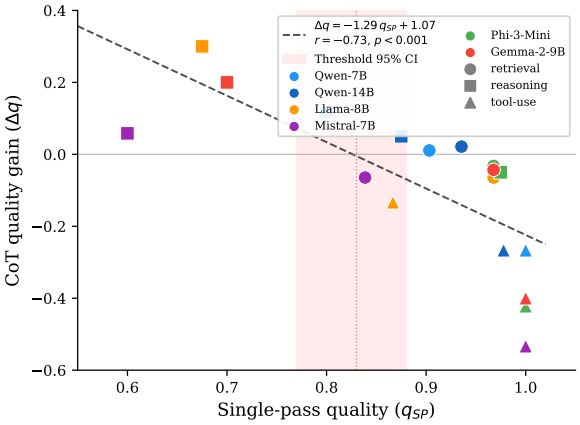

Figure 1: CoT quality gain vs. single-pass quality for all 18 model–family cells. The regression line ($\Delta q = -1.29\, q_{\text{SP}} + 1.07$, $r = -0.73$) crosses zero near $q_{\text{SP}} = 0.83$: above that, CoT hurts on average. Points colored by model, shaped by task family.

Table 2: Per-model aggregates, sorted by SP quality. Models with negative overall CoT gain ($*$) cluster by SP competence, not size.

Model	Params	SP q	CoT q	Δq	E_{m}
Mistral-7B	7B	.756	**.667***	**−.089**	3×
Llama-8B	8B	.814	.907	+.093	20×
Gemma-9B	9B	.849	.857	+.008	16×
Qwen-7B	7B	.872	.884	+.012	33×
Qwen-14B	14B	.915	**.899***	**−.016**	34×
Phi-3	3.8B	.977	**.868***	**−.109**	3×

The linear fit explains 53% of variance. The zero-crossing at $q_{\text{SP}} \approx 0.83$ acts as a natural decision boundary: above it, CoT is expected to produce zero or negative quality gain. We ran 10,000 bootstrap iterations and obtained a 95% CI of $[0.76, 0.88]$ for this crossing point; leave-one-out analysis yields a tighter range of $[0.82, 0.86]$ with std $= 0.007$. A logistic model gives essentially the same midpoint with no AIC improvement, so we report the simpler linear fit. Every tool-use cell falls in the regression zone (SP ≥ 0.867). The four cells with clearly positive CoT gain (Qwen-7B reasoning, Qwen-14B reasoning, Llama-8B reasoning, and Gemma-2-9B reasoning) all have SP quality at or below 0.875.

Table 2 rolls this up per model. Three of six models show net negative CoT gain. Sorting by single-pass quality (not parameter count) recovers a monotonic ordering.

An apparent exception: Mistral-7B has the lowest SP quality (0.756) but still shows negative overall CoT gain. This is because its tool-use regression is severe (−53.3 pp), dragging down the aggregate

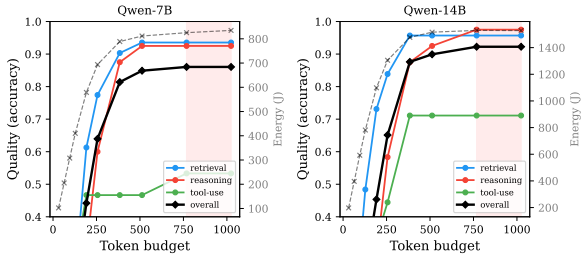

Figure 2: Quality (solid) and energy (dashed) vs. token budget for Qwen-7B (left) and Qwen-14B (right). Red shading: waste zone past quality saturation. 2,580 runs per model.

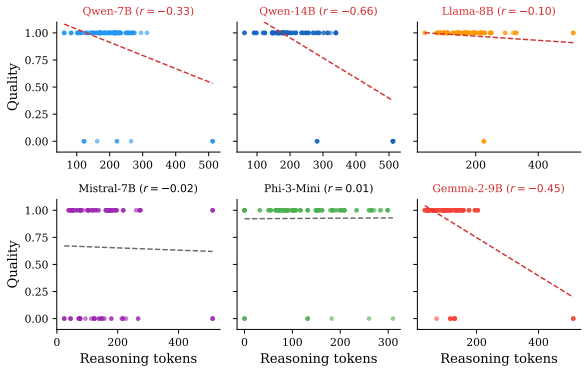

Figure 3: Reasoning tokens vs. quality under CoT for all six models (reasoning tasks only). Pearson r per panel. Red titles: negative correlations.

despite a positive reasoning gain (+5.8 pp). Looking at reasoning alone, Mistral's low SP quality (0.600) does predict the largest potential for improvement, consistent with the headroom story.

4.2 Token Budget Saturation

How much of the reasoning chain actually matters? We swept the token budget from 32 to 1,024 on both Qwen models (Figure 2). Quality peaks at budget 768 for both. For Qwen-14B, nearly all quality gain happens in the 128 → 384 range (95% of peak); the tokens beyond that, which represent 40% of the budget window, add less than 5% of quality while burning 34% of total energy.

Breaking this down by task family: retrieval saturates at budget 384 for both models, reasoning saturates at 512 (7B) and 768 (14B). A task-aware budget allocation (384 tokens for retrieval, 768 for reasoning, no CoT for tool-use) would capture over 95% of peak quality at a fraction of the energy.

4.3 More Tokens, Worse Answers

Figure 3 shows the per-model relationship between reasoning token count and quality on reasoning

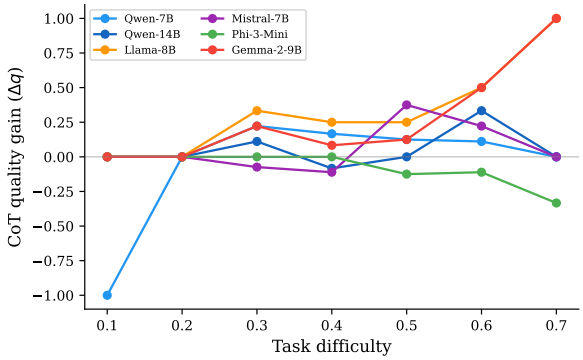

Figure 4: CoT gain by task difficulty on reasoning tasks. Low-SP models (Llama-8B, Gemma-9B) benefit across difficulty levels. High-SP Phi-3 shows flat or negative gain even on hard tasks.

tasks. Four of six models show a negative correlation, strongest for Qwen-14B ($r = -0.66$) and Gemma-2-9B ($r = -0.45$). The other two (Phi-3 and Mistral) are essentially flat.

Why would more tokens predict worse answers? There are at least three explanations, and our data cannot cleanly separate them: (a) harder tasks elicit both more tokens and more errors, making difficulty a confound; (b) models generate more tokens when uncertain, and uncertainty predicts failure; (c) extended reasoning introduces new failure modes like second-guessing or format drift. We report this as an empirical regularity rather than a causal claim.

4.4 Difficulty Does Not Override Headroom

One might expect harder tasks to always benefit from CoT, regardless of the model. Figure 4 says otherwise. Llama-8B and Gemma-2-9B (low SP quality) do show increasing CoT gains on harder tasks. But Phi-3-Mini (highest SP quality) shows negative CoT gain *at every difficulty level*. Even on the hardest reasoning tasks, Phi-3's single-pass competence is high enough that reasoning tokens add noise rather than signal. The headroom principle operates at a level above task difficulty.

4.5 Failure Analysis

To understand *how* CoT fails, we examined all 175 cases across every model where single-pass got the answer right and CoT got it wrong (Table 3).

The 175 regressions cluster around three patterns.

Tool-use format disruption dominates. 92 of the 175 regressions (54%) occur on tool-use tasks.

Table 3: CoT regressions (SP correct, CoT wrong) and improvements across all six models. Tool-use accounts for 54% of regressions.

Model	Family	SP=1	Regr	Impr	Net
Qwen-7B	Retr	84	3	4	+1
	Reas	96	6	20	+14
	Tool	45	12	0	−12
Qwen-14B	Retr	87	1	3	+2
	Reas	105	6	12	+6
	Tool	44	13	1	−12
Llama-8B	Retr	90	6	0	−6
	Reas	81	3	39	+36
	Tool	39	6	0	−6
Mistral-7B	Retr	78	15	9	−6
	Reas	81	22	20	−2
	Tool	45	24	0	−24
Phi-3	Retr	90	3	0	−3
	Reas	117	6	0	−6
	Tool	45	19	0	−19
Gemma-9B	Retr	90	6	2	−4
	Reas	84	6	30	+24
	Tool	45	18	0	−18
Total		**1,346**	**175**	**140**	**−35**

The mechanism is consistent: when asked to "think step by step" about a task requiring exact syntax (a Fibonacci function, an SQL query, a Docker command), the reasoning chain corrupts the output format. This is not overthinking. It is incompatibility between free-form reasoning and constrained generation. Tool-use improvements are essentially zero (1 total), confirming that CoT and structured output do not mix.

Reasoning improvements track headroom. On reasoning tasks, the three models with low SP quality show strong positive nets: Llama-8B (+36), Gemma-9B (+24), Qwen-7B (+14). The three with high SP quality show zero or negative: Phi-3 (−6), Mistral (−2), Qwen-14B (+6). The headroom principle holds even at the level of individual regression and improvement counts.

Retrieval regressions are rare. Only 34 of 175 regressions (19%) involve retrieval. CoT on retrieval is not actively harmful; it is simply expensive for negligible benefit.

5 Discussion

5.1 Headroom, Not Size

The most consistent predictor of CoT utility in our data is single-pass quality ($r = -0.73$), not parameter count. This resolves several results that a

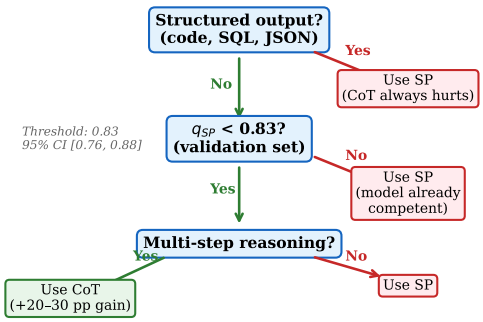

Figure 5: Decision flowchart for CoT routing. The 0.83 threshold (95% CI: $[0.76, 0.88]$) comes from the zero-crossing in Figure 1.

size-based story cannot explain: (1) Phi-3-Mini at 3.8B has worse CoT gain than Qwen-14B at 14B; (2) Llama-8B and Qwen-7B sit at similar sizes but opposite ends of the CoT benefit spectrum ($+9.3$ vs. $+1.2$ pp); (3) Gemma-2-9B benefits from CoT while smaller Mistral-7B does not. In each case, single-pass quality predicts what size cannot.

For practitioners, CoT routing decisions should be conditioned on estimated single-pass competence for the task at hand, not on model identity or scale.

5.2 A Simple Routing Heuristic

The data points toward a three-question rule (Figure 5). First, is the output structured (code, SQL, JSON)? If so, skip CoT. Second, is SP accuracy below $\sim$0.83 on a held-out validation set? If not, CoT is unlikely to help. Third, does the task involve multi-step reasoning? When all three conditions hold, CoT delivers 20–30 pp quality gains. When any fails, CoT is waste or worse.

A practical caveat: applying this rule requires estimating single-pass quality for the target task distribution. Whether the headroom principle holds at the individual query level, rather than the model–family level at which we measure it, remains open.

6 Conclusion

We measured 8,421 inference runs on six models from five families and found that single-pass quality, not model size and not model family, predicts whether chain-of-thought helps or hurts ($r = -0.73$). Three of six models produce worse answers with CoT than without it. CoT degrades tool-use quality for every model tested. On reasoning tasks, where CoT earns its keep, the energy costs run 2–61$\times$ over single-pass, and more than a third of that energy feeds tokens past diminishing returns. These findings argue for a shift from "always think" toward headroom-aware routing: let the model think when it needs to, and skip the reasoning when it already knows the answer.

Limitations

Our study has several limitations that should be considered when interpreting the results.

Scale of models tested. All six models fall in the 3.8B–14B parameter range. We do not test models above 14B, and the headroom relationship ($r = -0.73$, $N = 18$) could behave differently for larger models (e.g., 70B+) where CoT may interact differently with parametric knowledge. The correlation could also be non-linear at quality extremes not represented in our data.

Task coverage. Our benchmark (86 tasks, 3 families) excludes several important domains: code generation beyond simple functions, creative writing, long-context reasoning, mathematical proof, and multi-turn dialogue. The headroom principle may not generalize to these settings, particularly for tasks where the quality metric is not binary exact-match.

Decoding strategy. All runs use greedy decoding (temperature 0). Stochastic sampling, best-of-n, or majority voting could shift the token-quality relationship. In particular, self-consistency (Wang et al., 2023) is known to improve CoT quality through sampling diversity, which our setup does not capture.

Causal claims. The negative token-quality correlations (Section 4.3) are observational. We cannot determine whether longer reasoning chains *cause* worse answers or whether both are effects of task difficulty. Our failure analysis provides some mechanistic evidence, but a causal account would require controlled interventions on chain length.

Energy scope. Our measurements capture GPU energy consumption only via NVML counters. Total system energy (including CPU, memory, networking, cooling, and infrastructure overhead) is not measured. The energy multipliers we report are therefore lower bounds on the true cost ratio between CoT and single-pass inference.

Threshold generalization. The 0.83 threshold is derived from aggregate model–family cells, not

individual queries. Whether it applies at the per-query level in a production routing system, and whether it transfers across task distributions not in our benchmark, requires further investigation.

References

Anonymous. 2026. Towards green ai: The energy cost of llm babbling. *arXiv preprint arXiv:2602.05712*.

Damani and Collaborators. 2025. Learning how hard to think: Input-adaptive allocation of lm computation. *ICLR*.

DeepSeek AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. *ICML*.

Fernandez, Luccioni, and Strubell. 2025. Energy measurement methodology for llm inference. *Proceedings of ACL*.

Madaan and Collaborators. 2023. Self-refine: Iterative refinement with self-feedback. *NeurIPS*.

OpenAI. 2024. Learning to reason with llms. *OpenAI Blog*.

Singhal and Collaborators. 2024. The length bias in post-training: How rlhf and dpo produce verbose outputs. *arXiv preprint*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *ICLR*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Yonatan Bisk, and 1 others. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *ICLR*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and 1 others. 2023. Self-consistency improves chain of thought reasoning in language models. *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Yang and Collaborators. 2023. The hidden sampling gap in nvml power monitoring on a100. *arXiv preprint arXiv:2312.02741*.

Matei Zaharia and Collaborators. 2024. The shift from models to compound ai systems. *BAIR Blog*.

Zhou and Collaborators. 2026. Select-then-solve: Routing inference paradigms for llm reasoning. *arXiv preprint arXiv:2604.06753*.