

Shorthand for Thought: Compressing LLM Reasoning via Entropy-Guided Supertokens

Zhenyu Zhao

Lila Sciences*

zhenyuzhao@lila.ai

Sander Land

Writer, Inc.

sander@writer.com

Dan Bikel

Writer, Inc.

dan@writer.com

Waseem Alshikh

Writer, Inc.

waseem@writer.com

Abstract

Reasoning in Large Language Models incurs significant inference-time compute, yet the token-level information structure of reasoning traces remains underexplored. We observe that reasoning tokens split into two functional types: low-entropy *structural* tokens (recurring phrases that scaffold the reasoning process) and higher-entropy *organic* tokens (problem-specific content that drives toward a solution). This asymmetry motivates a simple, model-agnostic compression pipeline: apply cross-word BPE merges on a model’s own reasoning traces to derive *supertokens* that capture frequent structural patterns, then teach the model to adopt them via supervised fine-tuning. Across three model families and five mathematical reasoning benchmarks, our approach shortens reasoning traces by 8.1% on average; under a TOST equivalence analysis at a ± 2 pp margin, accuracy is equivalent or inconclusive on 13/15 model–benchmark cells (2 pass equivalence, 11 inconclusive, predominantly AIME at $N=30$), with non-equivalent degradation on 2/15 cells (DeepSeek-R1-Distill-Llama-70B on MATH-500 and OlympiadBench). Beyond compression, learned supertokens often align with interpretable reasoning moves such as backtracking, verification, and strategy shifts. This enables a compact structural analysis of reasoning traces: correct traces show more recovery and verification patterns, while incorrect traces show more repeated hedging and unresolved counterarguments. We release the full pipeline as open-source code.¹

1 Introduction

Large Language Models (LLMs) have achieved strong performance gains through chain-of-thought (CoT) reasoning, generating extended thinking tokens before producing answers (Wei et al., 2022; OpenAI, 2024; Guo et al., 2025). CoT reasoning has shown consistent benefits across mathematics, coding, and agentic planning (Lewkowycz et al., 2022; Chen et al., 2021; Yao et al., 2023).

However, reasoning traces incur substantial computational cost, and efforts to compress them have focused on replacing explicit reasoning with latent states or reducing length through distillation (Deng et al., 2024; Chen et al., 2024). Comparatively little attention has been given to the *content* of reasoning outputs: the linguistic structure and information density of the traces themselves.

In this paper, we take a first step toward modeling LLM reasoning in a more structured and principled manner. We propose that reasoning traces are not a homogeneous stream of tokens, but rather a composition of two functionally distinct token populations that we term **structural** and **organic** reasoning tokens:

- **Structural tokens:** recurring, formulaic phrases that scaffold reasoning, such as backtracking cues (“Wait, hold on”), verification phrases (“let us verify”), and

*Work done at Writer, Inc.

¹Code available at <https://github.com/Writer/shorthand-for-thought>.

strategy shifts (“Let me try”), which organize the reasoning flow but carry little problem-specific information.

- **Organic tokens:** problem-specific content that drives toward a solution, including mathematical expressions, variable bindings, intermediate results, and reasoning direction changes.

We formalize this distinction using information theory. By measuring per-token conditional entropy across a large corpus of CoT traces, we find that structural tokens concentrate in frequent multi-token patterns whose continuations are near-deterministic once the phrase begins, while organic tokens exhibit substantially higher entropy. We are not the first to observe this split: Wang et al. (2026) show that roughly 80% of CoT tokens are generated with high confidence, and Yuan et al. (2025) show that many thinking tokens contribute minimally to the final answer. Prior work uses the split to decide which tokens receive RL gradient or to prune low-importance content, in both cases holding the vocabulary fixed; we instead hold training fixed and change the vocabulary, turning the recurring low-entropy structural spans, whose information content is not commensurate with the cost of generating them one token at a time, into supertokens for lossless compression.

We exploit this by applying cross-word BPE merges (Liu et al., 2025) to each model’s own reasoning traces, deriving model-specific *supertokens* that we append to the base vocabulary and teach the model to adopt via supervised fine-tuning (SFT) of only the embedding layer, LM head, and a few transformer layers. The pipeline is model-agnostic: (1) collect reasoning traces, (2) derive supertokens via BPE merges, (3) fine-tune to adopt them.

This shortens reasoning traces by 8.1% on average across five benchmarks; a TOST equivalence analysis at ± 2 pp (Appendix I) finds accuracy equivalent or inconclusive on 13/15 model-benchmark cells (2 equivalent, 11 inconclusive, predominantly AIME at $N=30$), with non-equivalent degradation on 2/15 cells (DeepSeek-R1-Distill-Llama-70B on MATH-500 and OlympiadBench; Appendix A). Unlike content-level methods that prune steps or replace them with latent states, our vocabulary-level approach preserves the *complete, readable trace*: we compress the *expression* of reasoning, not its *content*. Because many supertokens correspond to recognizable reasoning moves, the compressed trace reads as a sequence of structural signposts (Figure 3).

Our contributions are:

1. A **simple, model-agnostic pipeline** for reasoning compression requiring only BPE-based supertoken derivation and supervised fine-tuning.
2. An **information-theoretic analysis** of reasoning traces via the structural/organic decomposition, showing that supertokens predominantly capture low-entropy structural patterns.
3. **8.1% reasoning compression with accuracy equivalent or inconclusive at ± 2 pp on 13/15 model-benchmark cells** (TOST; 11/15 inconclusive, predominantly AIME; 2/15 non-equivalent on DS-R1-Llama / MATH-500 and OlympiadBench) across three model families and five benchmarks, with evidence that learned supertokens align with interpretable reasoning idioms, enabling a descriptive structural analysis of correct and incorrect reasoning traces.

2 Related Work and Preliminaries

Superword tokenization. Standard BPE tokenizers (Sennrich et al., 2016; Radford et al., 2019) forbid tokens from spanning whitespace, limiting efficiency at larger vocabulary sizes. SuperBPE (Liu et al., 2025) and BoundlessBPE (Schmidt et al., 2025) relax this constraint to learn cross-whitespace superwords, cutting tokens by up to 33% on general text. Both target pretraining efficiency on general corpora; we instead apply the idea to reasoning traces from a specific model family, where high-frequency n-grams are structurally distinct and amenable to targeted filtering. As a drop-in tokenizer change, our method’s natural reference class is the tokenizer literature, where corpus-level gains between algorithms are

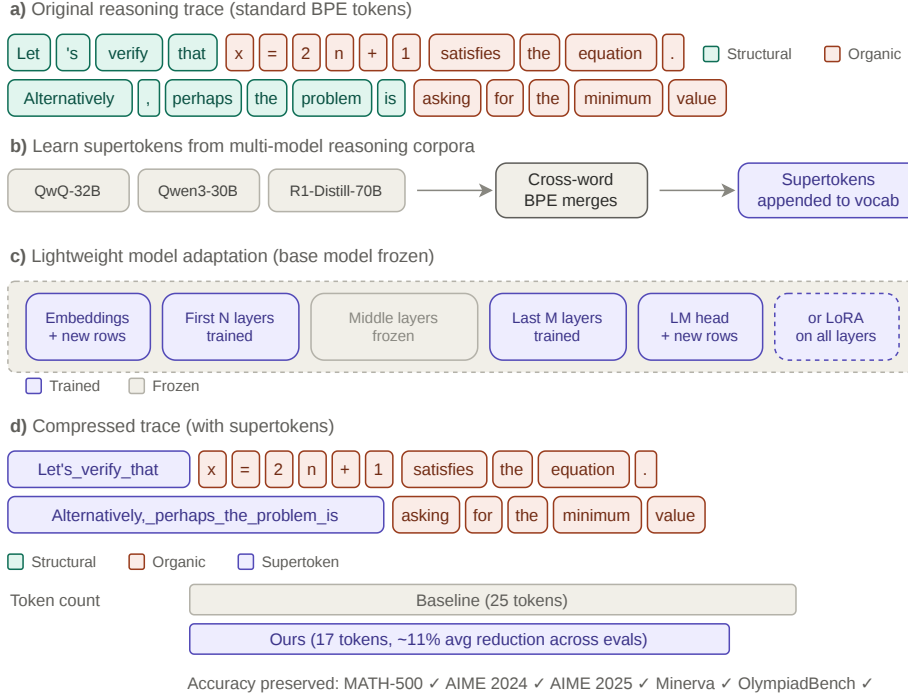


Figure 1: Overview of the pipeline: (1) derive reasoning-specific supertokens via BPE merges on reasoning traces, then (2) adapt the model to utilize these supertokens via supervised fine-tuning to produce shorter reasoning trajectories with accuracy equivalent or inconclusive at ± 2 pp on 13/15 model-benchmark cells under a TOST analysis (2 equivalent, 11 inconclusive; Appendix I).

typically 1–2% (PickyBPE (Chizhov et al., 2024): $\sim 1.1\%$), so our 8.1% is large; the 30–70% content-level figures (below) lie on a different axis and are not directly comparable.

Vocabulary expansion without retraining. Kaplan et al. (2025) show that LLMs internally assemble multi-token word representations into coherent single vectors by intermediate layers, enabling vocabulary expansion without retraining core parameters. We follow this approach, initializing new supertoken embeddings by averaging their constituent token embeddings.

Structured reasoning. Several works impose meta-level structure on the reasoning process (Yao et al., 2023; Besta et al., 2024; Yao et al., 2024); e.g., PLAN-AND-BUDGET (Lin et al., 2026) allocates a complexity-based budget over structured sub-questions and Tan et al. (2025) adds symbolic logic scaffolds. These modify the reasoning *process*; we instead compress its *output* at the vocabulary level.

Interpreting reasoning traces. Recent work analyzes chain-of-thought as structured behavioral objects: Gandhi et al. (2025) study cognitive behaviors (verification, backtracking, subgoal setting, backward chaining), Bogdan et al. (2025) identify steps that disproportionately influence later reasoning, and Shahariar et al. (2025) model traces with hierarchical finite-state abstractions. Our taxonomy overlaps with these but is derived from frequent supertokens learned for compression rather than a manual annotation scheme.

Latent compression of reasoning traces. A separate line replaces discrete tokens with latent representations: Token Assorted (Su et al., 2025) uses VQ-VAE latent tokens for initial steps and Coconut (Hao et al., 2025) reasons in continuous latent space. These achieve strong compression but require specialized training objectives and lose interpretability.

Method	Approach	Token ↓	ΔAcc	Interpret.
TokenSkip	Step pruning	30–40%	<4% drop	✗
R1-Compress	Chunk compression	~20%	0.6% drop	✗
ConMax	RL compression	43%	0.7% drop	✗
CtrlCoT	Dual-granularity	31%	+7.6 pp*	✗
Extra-CoT	Extreme ratio	>73%	+0.6 pp†	✗
Coconut	Latent space	High	Varies	✗
Token Assorted	VQ-VAE latent	High	Varies	✗
Ours	Vocabulary	8.1%	~0 pp‡	✓

Table 1: Comparison with CoT compression methods. Content-level and latent methods achieve higher compression but alter or remove reasoning steps. Our vocabulary-level approach preserves the complete readable trace, unlike methods that remove, summarize, or hide reasoning steps. *CtrlCoT’s +7.6 pp is relative to the TokenSkip baseline on MATH-500 (not the uncompressed model). †Extra-CoT’s +0.6 pp is relative to the uncompressed model on MATH-500 with Qwen3-1.7B. ‡Pooled aggregate across 15 model–benchmark cells; per-cell breakdown via TOST at ± 2 pp in Table 13 (2/15 equivalent, 11/15 inconclusive, 2/15 non-equivalent on DeepSeek-R1-Distill-Llama-70B / MATH-500 and OlympiadBench).

Explicit CoT compression. A growing body of work compresses chain-of-thought at the content level. Token- and chunk-level pruning or rewriting methods, including TokenSkip (Xia et al., 2025), R1-Compress (Wang et al., 2025), ConMax (Hu et al., 2026), CtrlCoT (Fan et al., 2026), and Extra-CoT (Tang et al., 2026), report 20–73% reductions, but largely on short-CoT, instruction-tuned backbones (GSM8K, MATH-500); on long-CoT traces the accuracy cost grows sharply, with independent re-evaluations showing TokenSkip incurring >20 pp drops at comparable compression (Hu et al., 2026; Fan et al., 2026). A related family controls length through training rather than per-token pruning: O1-Pruner (Luo et al., 2026), CoT-Valve (Ma et al., 2025), and AutoL2S (Luo et al., 2025) change *what* the model produces and rely on RL, a learned controller, or a switching mechanism. All operate at the content level: they remove, shorten, or hide explicit reasoning steps, creating a genuine efficiency–interpretability tradeoff, and are typically tuned in-distribution per benchmark. Our approach instead makes a one-time, out-of-distribution vocabulary change applied as a drop-in tokenizer, leaving the complete trace readable and grammatical (which matters for debugging and oversight), and we evaluate on long-CoT reasoning models (QwQ-32B, Qwen3-30B-A3B, DeepSeek-R1-Distill-Llama-70B) over AIME’24, AIME’25, MATH-500, Minerva, and Olympiad, where baseline traces span thousands of tokens. Table 1 positions our approach relative to these and the latent methods above.

3 Information-Theoretic Motivation

We now formalize the compression opportunity suggested by the structural/organic distinction, drawing on Shannon’s information-theoretic framework (Shannon, 1948; 1951). The key observation is that structural tokens concentrate in frequent *multi-token patterns* whose continuation tokens are highly predictable once the phrase begins: once a model commits to “Let me reconsider”, the remaining tokens of the phrase carry almost no information. We quantify this sequential redundancy via conditional entropy. Given a reasoning trace $T = (t_1, \dots, t_n)$ generated by a model with parameters θ , the conditional entropy at position i is:

$$H_\theta(t_i \mid t_{<i}) = - \sum_{v \in \mathcal{V}} P_\theta(t_i = v \mid t_{<i}) \log P_\theta(t_i = v \mid t_{<i}) \quad (1)$$

Let $\mathcal{M} \subset \{1, \dots, n\}$ denote positions within a multi-token merge span, with mean entropy $h_\mathcal{M} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} H_\theta(t_i \mid t_{<i})$. If a fraction $\rho = |\mathcal{M}|/n$ of tokens participate in merges, the theoretical compression ceiling is:

$$\Delta = \rho \cdot \left(1 - \frac{h_\mathcal{M}}{\log_2 |\mathcal{V}|} \right) \quad (2)$$

When $h_{\mathcal{M}} \ll \log_2 |\mathcal{V}|$, this simplifies to $\Delta \approx \rho$: nearly all tokens within merge spans can be absorbed, so a tokenization algorithm learning frequent patterns over reasoning traces should preferentially discover structural sequences.

4 Learning supertokens from reasoning traces

4.1 Vocabulary construction via SuperBPE

We apply SuperBPE (Liu et al., 2025) to a corpus of reasoning traces to derive a small set of *reasoning supertokens* that are appended to the base vocabulary of a pretrained LLM. The key design choices are (1) which token n-grams to consider as supertoken candidates, and (2) how to initialize the new token embeddings without retraining the transformer body.

Corpus and n-gram extraction. We generate reasoning traces from each target model on OpenThoughts3 (Guha et al., 2026) prompts; each model family receives its own supertokens from its own traces. The entropy analysis in Section 4.2 uses QwQ-32B (Qwen Team, 2025) as the representative model; cross-model generality is validated in Appendix F.2. N-gram counts are computed over reasoning traces only, with per-sample counts capped at 10 to prevent high-repetition samples from dominating merge selection (Land & Bartolo, 2024).

Filter design. Unfiltered BPE on reasoning text produces merges spanning syntactic boundaries (punctuation-initial sequences, cross-clause concatenations, multi-digit runs) unlikely to generalize. We define a *structural filter* restricting merges to four surface patterns: capitalized phrase-initial spans (e.g. “The answer is”), punctuation-plus-newline continuations, comma-led continuations, and space-prefixed single digits (multi-digit concatenations excluded). This reduces raw compression from $\sim 28\%$ to $\sim 10\%$ at 1000 merges (Figure 6, Appendix), producing semantically coherent reasoning phrases (Table 6, Appendix).

The four rules are a fixed surface-form predicate with no numeric hyperparameters and were specified independently of any downstream evaluation. To confirm they generalize beyond the trace slice used to inspect candidate n-grams, we set aside the last 20 of 120 OpenThoughts3-1.2M shards (200,000 traces, 1.01B reasoning tokens) as a held-out slice and re-derived filtered BPE merges on this slice using the identical filter. The held-out-derived vocabulary (954 merges) saves 7.64% of held-out reasoning tokens at 1000 merges, indistinguishable from the production vocabulary (757 merges) at 7.36% on the same data; 80.6% of the production top-100 merged phrases also appear in the held-out top-100. No retraining or new downstream evaluations were performed for this check.

Merge budget. After filtering, we retain the top- $K=250$ merges ranked by corpus frequency, each introducing one new vocabulary entry (e.g., $151,669 \rightarrow 151,919$ for the Qwen-family models). Table 6 and Figure 6 (Appendix) show representative statistics.

Embedding initialization. New token embeddings are initialized by averaging the embeddings of their constituent tokens, providing a stable starting point that requires no additional forward passes. This mean-of-constituents initialization is important in practice: a random initialization of the new supertoken rows under the same training budget failed to train stably and did not converge to usable supertoken adoption, so we did not pursue it further. We report this qualitatively rather than as a controlled ablation.

Tokenizer application. At both training and inference time, the base tokenizer is applied first, and supertoken merges are applied as a post-processing step over the resulting token sequence. This keeps the extended tokenizer compatible with the original model and avoids modifying tokenizer internals.

4.2 Entropy Validation of Supertokens

We now test the prediction from Section 3: do the merges discovered by our supertokenization algorithm capture sequential redundancy in reasoning traces? We compute per-token

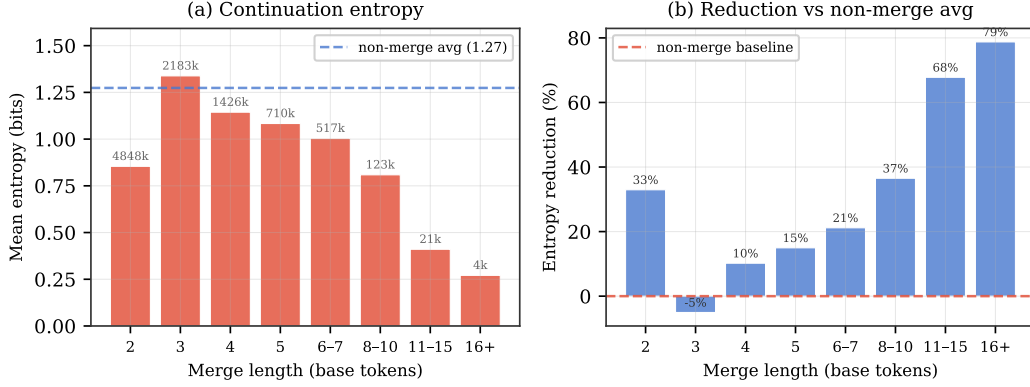


Figure 2: Continuation-token entropy by supertoken merge length. (a) Mean conditional entropy per bin (dashed line: non-merge average, 1.27 bits). The gap widens from 33% at length 2 to 79% at length 16+; length-3 merges are an outlier (−5%). (b) Relative entropy reduction ($1 - \bar{H}_{\text{cont}} / \bar{H}_{\text{non-merge}}$); longer merges capture progressively more deterministic continuations.

conditional entropy across 10,000 traces (115M tokens) from OpenThoughts3-1.2M using QwQ-32B as both generator and scorer, align each base token with the supertoken-based tokenization via character offsets, and classify tokens into three roles: **non-merged**, **first** (the opening token of a merge span), and **continuation** (subsequent tokens completing the phrase).

Merge length scaling. Our supertoken merges cover $\rho = 15.2\%$ of all token positions. As shown in Figure 2, continuation-token entropy decreases with merge length: length-2 merges already show a 33% entropy reduction relative to non-merge tokens, and merges of length ≥ 16 reach 79% reduction. This confirms that the merges target *sequential* redundancy, i.e., phrases where continuations become near-deterministic once the opening token is observed (see also Figure 7, Appendix). Length-3 merges are an exception, displaying entropy on par with non-merge tokens, likely because they contain a high concentration of transitioning tokens (Wang et al., 2026).

5 Experiments and Results

5.1 Experimental setup

Supertoken adoption training. We fine-tune each base model on OpenThoughts3 (Guha et al., 2026) with the extended tokenizer applied to assistant turns only. We compare three SFT modes: (i) *embedding-only* (new supertoken rows in `embed_tokens` and `lm_head`); (ii) *LoRA* (Hu et al., 2022) ($r=16$, $\alpha=32$ on attention projections plus embedding rows); and (iii) *layer-unfreezing* (first N and last M transformer layers plus embeddings). In all modes, gradient hooks zero out base-vocabulary rows so only supertoken embeddings receive updates. The layer-unfreezing configuration used for all main-table results ($N=3$ first, $M=1$ last transformer layers, with backbone learning rate 2×10^{-4}) is held fixed across all 15 model–benchmark cells; there is no per-cell hyperparameter tuning, and the LoRA-vs-unfreezing comparison (Table 8) is a training-intensity ablation rather than model selection. Exact values are listed in Appendix C.

Evaluation. We evaluate on five mathematical reasoning benchmarks (AIME’24, AIME’25, MATH-500 (Hendrycks et al., 2021), Minerva, OlympiadBench (He et al., 2024)) in a zero-shot setting using vLLM (Kwon et al., 2023) on 2 H200 GPUs, running each evaluation 5 times and reporting mean accuracy, completion tokens, and supertoken adoption rate (fraction of reasoning tokens from the expanded vocabulary). Throughout, we measure trace length in *tokens* rather than characters: each generated token is one decoder forward pass, so re-encoding the same reasoning in fewer tokens is a genuine reduction in inference

Benchmark	Baseline		SFT (no supertoken)		SFT (supertoken)			Improvement	
	Acc (%)	Avg. Tok. (↓)	Acc (%)	Avg. Tok. (↓)	Acc (%)	Avg. Tok. (↓)	SuperBPE (%)	ΔAcc (%)	ΔAvg. Tok. (%) (↓)
QWQ-32B									
AIME'24	77.50	14082	76.9	14580	77.7	13160	7.6	+0.2	-6.5
AIME'25	69.20	16651	68.3	16838	72.7	14880	7.7	+5.5	-10.6*
MATH-500	80.60	4441	81.0	4421	80.5	4183	10.7	-0.1	-5.8*
Minerva	34.0	5470	32.6	5695	33.5	5418	8.15	-1.5	-1.1
Olympiad	55.8	9584	56.7	9517	56.5	9256	11.7	+1.3	-3.0*
Average	63.4	10046	63.1	10210	64.2	9379	9.17	+1.3	-6.6*
QWEN-3-30B-A3B									
AIME'24	91.3	16305	91.9	16125	90.3	15246	10.7	-1.2	-6.5
AIME'25	84.0	19491	87.0	19275	84.1	18517	11.1	+0.1	-5.0
MATH-500	85.7	5488	85.8	5454	86.8	4849	12.9	+1.2	-11.6*
Minerva	37.4	3595	38.2	3566	37.0	3571	9.7	-1.1	-0.66
Olympiad	65.3	13309	64.5	13220	66.8	12372	11.7	+2.3	-7.0*
Average	74.7	11638	73.5	11528	73	10910	11.2	-1.3	-6.3*
DeepSeek-R1-Llama-70B-Distill									
AIME'24	66.7	9059	68.2	9037	70.2	7519	12.1	+5.1	-17.0*
AIME'25	54.5	11332	53.7	11329	53.4	9390	11.6	-2.0	-17.0*
MATH-500	78.1	2701	75.2	2683	74.8	2541	12.2	-3.1	-5.9*
Minerva	31.6	3239	32.5	3217	31.4	3126	9.54	-0.6	-3.5*
Olympiad	51.0	6373	50.3	6282	49.5	6370	11.4	-2.9	-0.1
Average	56.4	6541	56.0	6510	55.9	5749	11.4	-0.8	-11.5*

Table 2: SFT encourages supertoken adoption, reducing reasoning trace length across three model families and five math benchmarks. The **SFT (no supertoken)** columns report a control where each model is fine-tuned on OpenThoughts3 with its original tokenizer (no vocabulary expansion), isolating the contribution of supertoken adoption: this control leaves trace length essentially unchanged, so the compression is attributable to the supertokens rather than to SFT on the data. Accuracy deltas are evaluated by a TOST equivalence test at ± 2 pp (Appendix I, Table 13): 2/15 cells are equivalent, 11/15 inconclusive (predominantly AIME at $N=30$), and 2/15 (DS-R1-Llama on MATH-500 and OlympiadBench) fail equivalence with a real accuracy loss. * marks token reductions whose 95% CI excludes zero ($p < 0.05$).

compute, not merely a change in how length is counted. Wall-clock latency reductions generally track or exceed token reductions (Table 3).

5.2 Adoption of supertokens and compression of reasoning trajectories

As shown in Table 2, SFT on a retokenized dataset consistently drives supertoken adoption across all three model families and five benchmarks, reducing reasoning trace length by 6–12% on average. The reduction is attributable to the expanded vocabulary rather than to any behavioral shortening induced by fine-tuning: the *SFT (no supertoken)* control, trained on the same data with identical hyperparameters but no vocabulary expansion, leaves token length essentially unchanged (-0.47% on average, TOST-equivalent to zero), whereas adding supertokens yields the full -8.1% . The model therefore expresses the same reasoning content in fewer tokens rather than producing less reasoning, consistent with the lossless, content-preserving reading of the method. The two Qwen-family models (QwQ-32B (Qwen Team, 2025), Qwen3-30B-A3B (Yang et al., 2025)) show robust accuracy (≤ 1.3 pp average change); DeepSeek-R1-Distill-Llama-70B achieves the largest compression (up to -17% on AIME) but exhibits a -3.1 pp accuracy drop on MATH-500 whose mechanism we do not establish in this manuscript and discuss in Appendix A. We support the accuracy-preservation claim with a two one-sided test (TOST) at a single equivalence margin of ± 2 pp, using a paired t -test on per-question accuracy averaged across the 5 evaluation runs (Appendix I). Of the 15 model-benchmark cells, 2/15 demonstrate accuracy equivalence at ± 2 pp (QwQ-32B and Qwen3-30B-A3B on MATH-500), 11/15 are inconclusive (90% CI straddles a margin bound, predominantly AIME at $N=30$), and 2/15 fail equivalence with a real accuracy loss: DeepSeek-R1-Distill-Llama-70B on MATH-500 (-3.1 pp, 90% CI $[-5.18, -1.02]$) and on OlympiadBench (-2.9 pp, 90% CI $[-4.97, -0.83]$); both are discussed in Appendix A. Token reductions marked * in Table 2 are significant ($p < 0.05$), confirming that the compression gains are real.

Wall-clock latency. Table 3 reports wall-clock latency (mean seconds per sample) for all models and benchmarks. Evaluations are performed using vLLM on $2\times$ H200 GPUs with greedy decoding. Latency reductions generally track or exceed token count reductions, confirming that the shorter reasoning traces from supertoken adoption translate into practical inference speedups. Averaged across the three models, wall-clock latency drops by $\sim 10.5\%$ (QwQ-32B -14.1% , Qwen3-30B-A3B -13.1% , DeepSeek-R1-Distill-Llama-70B -4.4%). QwQ-32B shows the most consistent gains, while DeepSeek-R1-Distill-Llama-70B shows variable results across benchmarks.

Benchmark	Baseline Lat. (s)	SFT Lat. (s)	Δ Lat. (%) ($\downarrow$)
QWQ-32B			
AIME'24	507	466	-8.1
AIME'25	759	649	-14
MATH-500	1538	1252	-18.9
Minerva	620	545	-12
Olympiad	3416	2872	-16
Average	1246	1157	-14.1
QWEN-3-30B-A3B			
AIME'24	2120	1844	-13.1
AIME'25	2453	1922	-21.6
MATH-500	2281	1882	-17.5
Minerva	3595	3571	-0.7
Olympiad	4833	4732	-2.1
Average	3056	2790	-13.1
DeepSeek-R1-Llama-70B-Distill			
AIME'24	618	619.3	+ 0.2
AIME'25	1620	1483	-8.5
MATH-500	2329	1989	-14.6
Minerva	823	831.2	+1.0
Olympiad	9110	8936	-1.9
Average	2900	2772	-4.4

Table 3: Wall-clock latency (mean seconds per sample) for baseline and SFT with supertokens, served via SGLang or VLLM on H200 GPUs. The Δ Lat. column reports the relative change from Baseline to SFT (supertoken). Latency reductions exceed or track token count reductions from Table 2.

5.3 Structural analysis of reasoning trajectories

Beyond compression, the learned supertokens provide a compact and interpretable view of recurring reasoning structure. We classify supertokens into the taxonomy shown in Table 4 using deterministic keyword-based rules applied in priority order to each merge’s decoded text (full rule specification in Appendix B.2). This maps each trace into a sequence of coarse reasoning moves. Although similar categories can be detected by applying string-matching rules directly to raw reasoning traces, the advantage of supertokens is that high-frequency recurring idioms are learned from the model’s own trace distribution and emitted as discrete vocabulary items.

Figure 3 shows two reasoning traces as illustrative examples, highlighting the logical flow and key transitions throughout each trajectory. The correct trace exhibits a balanced structure: the model explores strategies, verifies intermediate results, and backtracks sparingly.

Category	Description	Representative Supertokens
Backtracking	Self-correction; pausing to reconsider	Wait, hold on But wait Wait, no
Hedging	Expressing uncertainty or tentativeness	, maybe But maybe , so maybe
Verification	Checking or validating intermediate results	, right , correct let's check
Counterargument	Introducing a contradiction or counterpoint	, but , but the , but we But the
Strategy Shift	Changing approach or trying a new method	Let's Let me Now, let's Let's try
Problem Ref.	Re-reading or referencing the problem	, the problem says But the problem
Sequencing	Organizing steps or structural ordering	First, Similarly, Given that Now, the
Reasoning	Forward-flowing connectives	, so , and , which is Therefore,
Computation	Digits or single-letter variables	1 2 , x , n , k

Table 4: Supertoken structural taxonomy. Each merge in the SuperBPE vocabulary is classified into one of nine structural categories based on its decoded text. Representative supertokens are shown for each category.

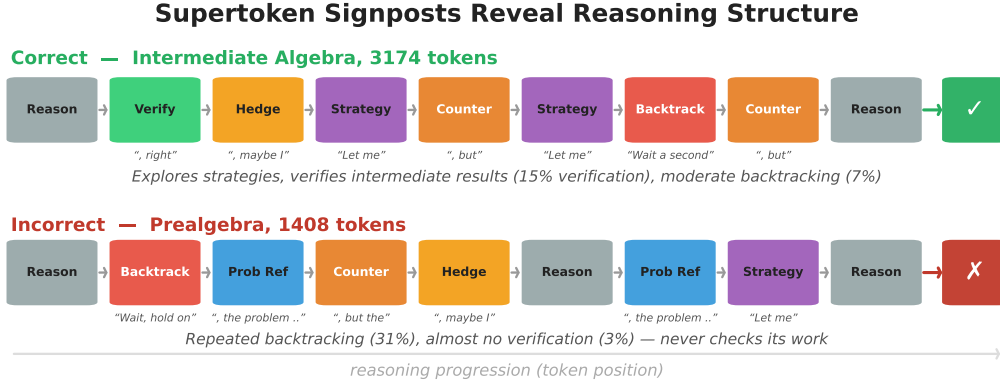


Figure 3: Supertoken signposts reveal reasoning structure. Each block represents a structural reasoning move detected via supertokens. The correct trace (top) shows diverse, balanced reasoning with verification; the incorrect trace (bottom) is dominated by repeated backtracking without verification.

In contrast, the incorrect trace is dominated by repeated backtracking with little verification, suggesting repeated self-correction without sufficient checking.

Aggregating supertoken transitions by answer correctness shows that the structural differences illustrated above extend across the full evaluation set. We report two clusters of transition bigrams that are over-represented in one class relative to the other (Table 5); these are aggregate over-representations across the evaluation pool, not held-out predictive evaluations of trace correctness. *Incorrect* traces show elevated rates of what we descriptively label **confusion cycles**: problem-reference→hedging ($2.1\times$ over-represented), counterargument→problem-reference ($2.0\times$), and hedging→hedging ($1.4\times$). *Correct* traces show elevated rates of **productive recovery**: problem-reference→strategy-shift ($3.0\times$), verification→strategy-shift ($2.1\times$), and reasoning→verification ($1.6\times$). The full transition probability matrix is shown in Appendix G.

From	To	Ratio	Interpretation
<i>Problematic transitions</i> (over-represented in incorrect traces)			
Sequencing	Sequencing	3.7×	Repeated reorganization without progress
Problem Ref.	Hedging	2.1×	Re-reads problem → becomes uncertain
Counterargument	Problem Ref.	2.0×	Contradiction → re-reads problem (confusion loop)
Hedging	Hedging	1.4×	Sustained unresolved uncertainty
Counterargument	Counterargument	1.3×	Contradictions accumulating unresolved
<i>Productive transitions</i> (over-represented in correct traces)			
Problem Ref.	Strategy Shift	3.0×	Re-reads problem → devises new plan
Verification	Strategy Shift	2.1×	Checks result → adapts approach
Reasoning	Verification	1.6×	Forward reasoning → verifies work
Backtracking	Strategy Shift	1.5×	Recognizes error → pivots constructively

Table 5: Bigram supertoken transitions over-represented by trace correctness. *Problematic* bigrams are over-represented in incorrect traces; *productive* bigrams are over-represented in correct traces. Ratios are aggregate over-representations across the evaluation pool; these are descriptive correlational patterns and not held-out predictive evaluations of correctness.

6 Discussion

Our results suggest that reasoning traces contain a substantial amount of repeated low-entropy scaffolding. By learning vocabulary items for these recurring phrases, supertokens compress the expression of reasoning without removing intermediate steps or replacing them with latent states. This gives a different trade-off from content-level or latent CoT compression methods: the compression ratio is more modest, but the resulting traces remain readable and largely faithful to the original reasoning process. The cost profile also differs: supertoken adoption requires only a single small round of plain SFT, with no reward model, controller, or inference-time routing, and the resulting tokenizer and model are a permanent drop-in change with zero per-request overhead, so the training cost is one-time and amortized over all subsequent inference.

A useful byproduct of this vocabulary-level compression is interpretability. Many learned supertokens correspond to recognizable reasoning idioms such as backtracking, verification, strategy shifts, and hedging, making the compressed trace easier to inspect at a structural level. These analyses do not require supertokens in principle; similar categories can be detected by string matching over raw traces, while supertokens provide a learned vocabulary-level interface for inspecting frequent idioms.

The transition patterns we observe are correlational: they describe aggregate over-representations across our evaluation pool, not whether any individual transition is predictive of correctness on a held-out trace. We therefore present them as a descriptive structural analysis only. Whether bigram-level transition features add held-out predictive signal beyond simple baselines such as trace length, and, if so, whether they could in turn serve as early-stopping or reward-shaping signals in RL-based reasoning training, is a natural follow-up that we leave to future work.

Overall, reasoning supertokens provide a simple, model-agnostic way to compress LLM reasoning traces while preserving readability, without pruning steps or replacing them with latent states. Across three model families and five benchmarks, this vocabulary-level approach achieves 8.1% average compression with accuracy equivalent or inconclusive at ± 2 pp on 13/15 model–benchmark cells (TOST; 2 equivalent, 11 inconclusive), with two non-equivalent degradations on DeepSeek-R1-Distill-Llama-70B (Appendix A), while exposing recurring structural patterns in the generated traces. Appendix A discusses the limitations of this work: the modest compression ratio relative to content-level methods, the restriction of our evaluation to mathematical reasoning, the dependence on representative traces from the target model, and the accuracy degradation observed on DeepSeek-R1-Distill-Llama-70B.

References

- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and et al. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar. 2024. doi: 10.1609/aaai.v38i16.29720. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29720>.
- Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter?, 2025. URL <https://arxiv.org/abs/2506.19143>.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. Learning to maximize mutual information for chain-of-thought distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6857–6868, 2024.
- Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. BPE gets picky: Efficient vocabulary refinement during tokenizer training. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16587–16604, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.925. URL <https://aclanthology.org/2024.emnlp-main.925/>.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=mZn2Xyh9Ec>.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit CoT to implicit CoT: Learning to internalize CoT step by step. *arXiv preprint arXiv:2405.14838*, 2024.
- Zhenxuan Fan, Jie Cao, Yang Dai, Zheqi Lv, Wenqiao Zhang, Zhongle Xie, Peng Lu, and Beng Chin Ooi. Ctrlcot: Dual-granularity chain-of-thought compression for controllable reasoning, 2026. URL <https://arxiv.org/abs/2601.20467>.
- Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STars. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=QGJ9ttXLty>.
- Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=7xjoTuaNmN>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui

- Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, Sept 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL <http://dx.doi.org/10.1038/s41586-025-09422-z>.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language model to reason in a continuous latent space, 2025. URL <https://openreview.net/forum?id=tG4SgayTtk>.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL <https://aclanthology.org/2024.acl-long.211/>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=7Bywt2mQsCe>.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Minda Hu, Zexuan Qiu, Zenan Xu, Kun Li, Bo Zhou, and Irwin King. Conmax: Confidence-maximizing compression for efficient chain-of-thought reasoning, 2026. URL <https://arxiv.org/abs/2601.04973>.
- Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=328vch6tRs>.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

- Sander Land and Max Bartolo. Fishing for Magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL <https://aclanthology.org/2024.emnlp-main.649/>.
- Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=IFXTZERXdM7>.
- Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=ctspw4CqBS>.
- Alisa Liu, Jonathan Hayase, Valentin Hofmann, Sewoong Oh, Noah A. Smith, and Yejin Choi. SuperBPE: Space travel for language models. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=lcDRvffeNP>.
- Feng Luo, Yu-Neng Chuang, Guanchu Wang, Hoang Anh Duy Le, Shaochen Zhong, Hongyi Liu, Jiayi Yuan, Yang Sui, Vladimir Braverman, Vipin Chaudhary, and Xia Hu. Autol2s: Auto long-short reasoning for efficient large language models, 2025. URL <https://arxiv.org/abs/2505.22662>.
- Haotian Luo, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Xiaochun Cao, Dacheng Tao, Naiqiang Tan, and Li Shen. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14242–14257, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.697. URL <https://aclanthology.org/2026.findings-acl.697/>.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. CoT-valve: Length-compressible chain-of-thought tuning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6025–6035, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.300. URL <https://aclanthology.org/2025.acl-long.300/>.
- OpenAI. Learning to reason with LLMs, 2024. Blog post, September 2024.
- Qwen Team. QwQ-32B: Embracing the power of reinforcement learning. <https://qwenlm.github.io/blog/qwq-32b/>, March 2025. Blog post.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.
- Craig W Schmidt, Varshini Reddy, Chris Tanner, and Yuval Pinter. Boundless byte pair encoding: Breaking the pre-tokenization barrier. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=oPAjXGV8qQ>.
- Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162/>.

- G M Shahariar, Ali Nazari, Erfan Shayegani, and Nael Abu-Ghazaleh. Modeling hierarchical thinking in large reasoning models, 2025. URL <https://arxiv.org/abs/2510.22437>.
- Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- Claude E Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64, 1951.
- Dijia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=hYfOPXrbUr>.
- Xingwei Tan, Marco Valentino, Mahmud Elahi Akhter, Maria Liakata, and Nikolaos Aletras. Enhancing logical reasoning in language models via symbolically-guided Monte Carlo process supervision. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31886–31900, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1624. URL <https://aclanthology.org/2025.emnlp-main.1624/>.
- Yuntian Tang, Bohan Jia, Wenxuan Huang, Lianyue Zhang, Jiao Xie, Wenxi Li, Wei Li, Jie Hu, Xinghao Chen, Rongrong Ji, and Shaohui Lin. Towards efficient large language reasoning models via extreme-ratio chain-of-thought compression. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=hwJ2SZtVij>.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL <https://openreview.net/forum?id=yfcpdY4gMP>.
- Yibo Wang, Haotian Luo, Huanjin Yao, Tiansheng Huang, Haiying He, Rui Liu, Naiqiang Tan, Jiaxing Huang, Xiaochun Cao, Dacheng Tao, and Li Shen. R1-compress: Long chain-of-thought compression via chunk compression and search, 2025. URL <https://arxiv.org/abs/2505.16838>.
- 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. TokenSkip: Controllable chain-of-thought compression in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.165. URL <https://aclanthology.org/2025.emnlp-main.165/>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Zhang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yao Yao, Zuchao Li, and Hai Zhao. GoT: Effective graph-of-thought reasoning in language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2901–2921, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.183. URL <https://aclanthology.org/2024.findings-naacl.183/>.

Hang Yuan, Bin Yu, Haotian Li, Shijun Yang, Christina Dan Wang, Zhou Yu, Xueyin Xu, Weizhen Qi, and Kai Chen. Not all tokens are what you need in thinking, 2025. URL <https://arxiv.org/abs/2505.17827>.

Jeffrey Zhou et al. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A Limitations

Our approach achieves more modest compression ($\sim 8.1\%$) compared to content-level CoT compression methods that report 30–70% reductions (Table 1). This comparison spans two different reference classes: as a drop-in tokenizer change, our method should be measured against the tokenizer-algorithm literature, where corpus-level gains are typically 1–2% (PickyBPE (Chizhov et al., 2024): $\sim 1.1\%$) and 8.1% is large, whereas the 30–70% figures come from content-level methods that remove or shorten reasoning steps and report measurable accuracy loss on harder benchmarks. The more modest ratio is a deliberate design choice: by operating at the vocabulary level, we compress the *tokenization* of reasoning rather than its *content*, preserving the complete human-readable trace.

Our method also depends on the availability of representative reasoning traces from the target model, and the learned vocabulary may not transfer cleanly across tokenizer families.

Domain coverage. Our evaluation is restricted to mathematical reasoning. The structural idioms our supertokens capture (e.g., “let me think”, “wait”, “verify”) are largely domain-general reasoning markers, so we expect partial transfer, but we have not validated this and do not claim it. Coding, agentic planning, and scientific or long-form QA are the untested domains and the natural next evaluation. The IFEval results in Appendix H indicate that general instruction-following is preserved after supertoken SFT, but this does not by itself establish that the compression transfers to non-math reasoning.

Accuracy degradation on DeepSeek-R1-Distill-Llama-70B. Our TOST analysis at ± 2 pp (Appendix I) identifies two model-benchmark cells where the accuracy drop is non-equivalent and the 90% CI excludes 0 in the negative direction (a real accuracy loss): DeepSeek-R1-Distill-Llama-70B on MATH-500 ($\Delta = -3.1$ pp, 90% CI $[-5.18, -1.02]$) and on OlympiadBench ($\Delta = -2.9$ pp, 90% CI $[-4.97, -0.83]$). Both failing cells are on the same model, on which we also observe the largest absolute token compression. A third DeepSeek-R1-Distill-Llama-70B cell with a negative point estimate, AIME’25 ($\Delta = -2.0$ pp), is statistically inconclusive due to limited sample size ($N=30$, 90% CI $[-12.11, +8.11]$). The Qwen-family models show no non-equivalent cells. Cleanly disentangling tokenizer-family effects from model-specific effects requires evaluating a second Llama-family model (e.g., R1-Distill-Llama-8B or Llama-3.3-70B-Instruct), which is the priority follow-up we plan to run.

B Reasoning Supertokens

Table 6 shows representative supertokens under different filtering rules. The *structural filter* (Section 4.1) removes high-frequency patterns unrelated to reasoning (e.g., “is the”, “in the”) and possessive artifacts starting with “’s”, while retaining compositional reasoning phrases such as “Let’s check” and “Let’s reconsider”.

Token	No filter	Filtered
<i>Kept by both strategies</i>		
. \n	261,553	261,553
. \n\n	254,138	254,138
: \n	243,970	243,970
, _the	242,352	242,352
_ 1	264,450	264,450
, _then	193,242	193,242
What _if	136,224	136,224
So _the	135,171	135,171
<i>Removed by filter</i>		
_is _the	241,551	–
_is _a	241,413	–
_in _the	231,043	–
_of _the	230,200	–
’s _check	219,883	–
’s _assume	165,624	–
impl ies	196,004	–
? _No . \n	210,179	–
<i>Surfaced by filter</i>		
Let ’s	–	189,651
Is _it	–	156,033
Let ’s _check	–	152,862
Is _it _possible	–	151,487
This _is	–	147,188
Is _there	–	142,627
Wait , _if	–	127,788
Let ’s _assume	–	124,009
This _implies	–	105,077
Let ’s _reconsider	–	47,570
Let ’s _consider	–	45,207
Frequencies at 150-merge budget. – indicates token absent from that merge table.		

Table 6: Representative super-tokens from unfiltered and filtered BPE. The filter removes syntactically promiscuous merges and math artifacts, and by admitting apostrophes surfaces a compositional family of reasoning-specific phrases built on `Let ’s` as a shared prefix.

B.1 Exemplary Reasoning Traces

Figures 4 and 5 show two complete reasoning traces rendered at the token level, with supertoken merge positions colored by their structural category (Table 4). Each figure

consists of a compact *ribbon* overview (top) showing the category distribution across the full trace, and three *zoomed windows* that reveal token-level detail at structurally interesting points.

The correct trace (Figure 4) exhibits a diverse mix of categories throughout: strategy shifts (purple) and verification (green) appear regularly, interspersed with forward reasoning. In contrast, the incorrect trace (Figure 5) is dominated by backtracking (red) and counterarguments (orange) with notably fewer verification events, consistent with the confusion-cycle pattern described in Section 5.3.

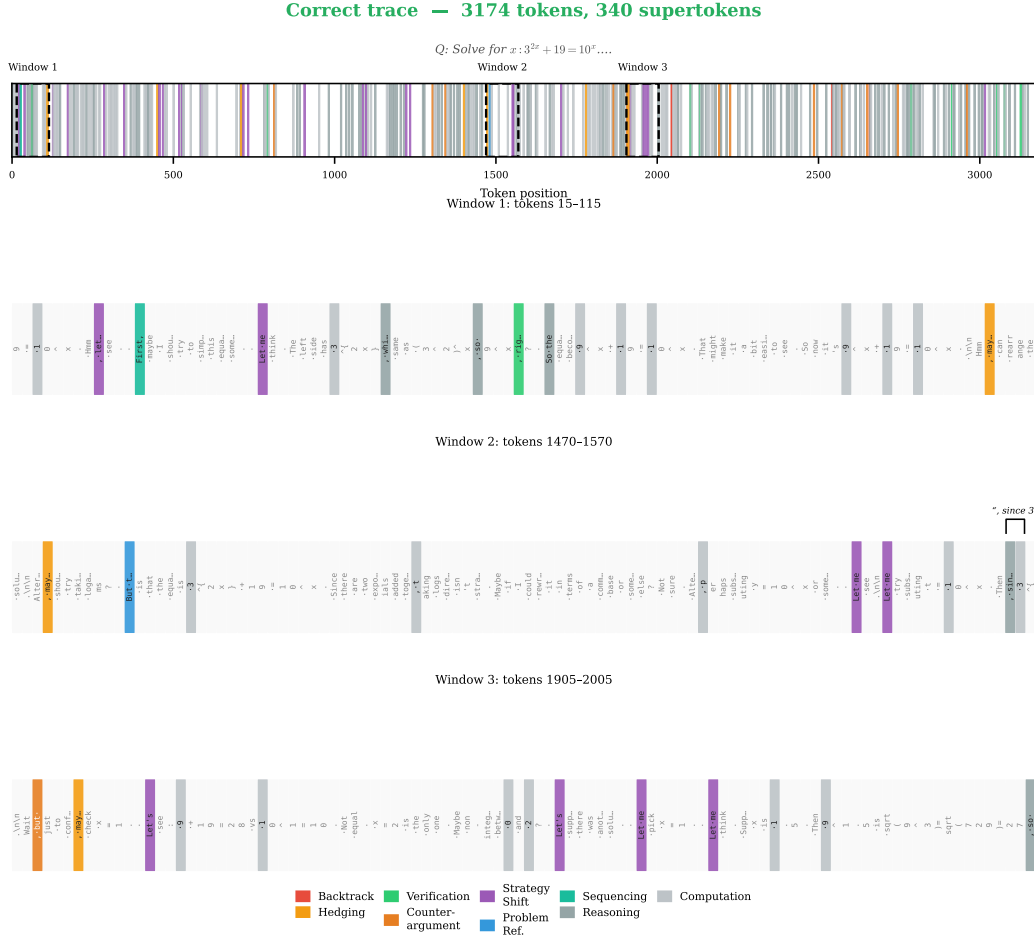


Figure 4: Correct reasoning trace with supertoken categories. Top: ribbon overview showing category distribution across all 3,174 tokens. Bottom: three zoomed windows at structurally interesting positions. Colors correspond to the taxonomy in Table 4; gray tokens are non-merged (organic content).

B.2 Taxonomy Construction Details

The nine-category taxonomy in Table 4 is applied via a deterministic, rule-based classifier that operates on the decoded text of each supertoken merge. Classification proceeds by evaluating pattern-matching rules in a fixed priority order; the first matching rule determines the category. No machine-learned component or manual per-token annotation is involved: the same script reproducibly assigns categories to any set of merges.

Rule summary. Table 7 lists the keyword or regex pattern that triggers each category, ordered by evaluation priority (highest first). Higher-priority rules ensure that composite

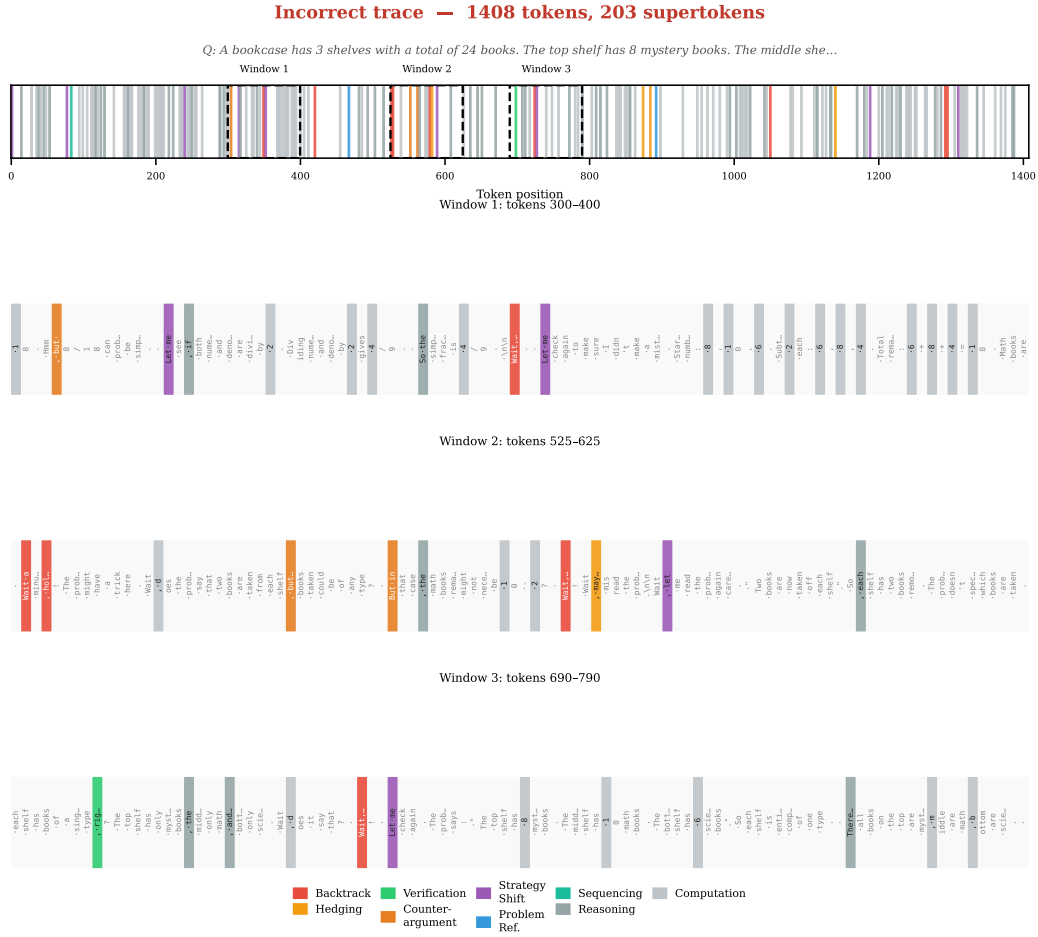


Figure 5: Incorrect reasoning trace with supertoken categories. The ribbon overview reveals a higher density of backtracking (red) and counterargument (orange) tokens compared to the correct trace. Zoomed windows show repeated self-correction cycles without verification, illustrating the confusion patterns identified in Section 5.3.

phrases are resolved correctly; for example, “Wait, let’s check” matches *Backtracking* (via the “wait” prefix) before it could match *Verification* (via “check”) or *Strategy Shift* (via “let’s”).

Pri.	Category	Key patterns (decoded text)	N
1	Backtracking	starts with “Wait”; contains “hold on”, “but actually”	16
2	Hedging	contains “maybe”	12
3	Verification	contains “check”, “matches”, “which is what”; exact match: “correct”, “yes”, “right”, “good”, “impossible”	12
4	Problem Ref.	contains “problem” or “it says”	8
5	Strategy Shift	starts with “Let’s”, “Let me”, “So let”, “But let”	12
6	Sequencing	starts with “First”, “Similarly”, “Given that”, “Now, the”	5
7	Computation	single digit (0–9) or single-letter variable (excl. “a” and “I” to avoid articles/pronouns)	36
8	Counterargument	starts with “But ” or “, but”	25
9	Reasoning	comma-led connectives (“, so”, “, and”, “, which”, ...); “Therefore”; “I think”; fallback: any “, ” + word	124

Table 7: Taxonomy classification rules applied in priority order. N = number of the 250 supertokens assigned to each category. All 250 merges are classified; no merges fall into an “other” residual category.

Coverage and ambiguity. Of the 250 supertokens, all are classified by the deterministic rules (0 fall into the residual “other” category). Because rules are evaluated in priority order, composite phrases that could match multiple categories (e.g., “, but maybe the problem” contains keywords for Counterargument, Hedging, and Problem Reference) are resolved deterministically. In this example, the “maybe” rule (priority 2) fires before “problem” (priority 4), yielding *Hedging*, reflecting that the dominant pragmatic function of the phrase is expressing uncertainty rather than referencing the problem. We verified the priority ordering by manually inspecting all 25 supertokens that match keywords from two or more categories; in each case the highest-priority rule assigns the pragmatically dominant function.

The full classification script and its output (a JSON mapping from token ID to category) are released with our code, enabling exact reproduction of all taxonomy-dependent analyses in §5.3.

C Experimental Setup

This section describes the compute environment, hyperparameters, and ablation studies used in our experiments.

C.1 Compute and infrastructure

All SFT runs use $8 \times \text{H200}$ GPUs. QwQ-32B and Qwen3-30B-A3B each train for 1 epoch on OpenThoughts3 ($\sim 114\text{K}$ samples); DeepSeek-R1-Distill-Llama-70B uses the same data and epoch count. Total wall-clock training time ranges from 4–12 hours depending on model size.

C.2 Training hyperparameters

Training uses a per-device batch size of 1 with 16 gradient accumulation steps (effective batch size 128), a learning rate of 2×10^{-4} with cosine decay and 3% linear warmup, for 1 epoch. For the partial layer-unfreezing mode used in all main-table runs, we unfreeze the first $N=3$ and last $M=1$ transformer layers together with the `embed_tokens` and `lm_head` rows; these values, the learning rate, and the warmup schedule are held fixed across all three models and all five benchmarks, with no per-cell tuning. We use bfloat16 mixed precision, Flash Attention 2 (Dao, 2024), gradient checkpointing, and Fully Sharded Data Parallel (FSDP) with full sharding. Checkpoints are saved every 100 steps. A separate learning rate can optionally be applied to the embedding/`lm_head` parameters independently from the backbone learning rate.

C.3 Minimum training ablation

To understand the minimum training needed for supertoken adoption, we tested two reduced configurations on Qwen3-30B-A3B: (i) masking losses from all tokens except merged supertokens; and (ii) LoRA with ranks $r \in \{1, 2, 4, 8, 16\}$. Results for the LoRA rank sweep are presented in Section D.1.

D Training Ablations

Here we present the results of the three different training approaches we tested: Embedding-only, LoRA, and partial-layer unfreezing. All three methods freeze the base model’s pre-trained weights and differ only in which additional parameters receive gradient updates during SFT. Embedding-only trains just the new supertoken rows in `embed_tokens` and `lm_head`. LoRA adds low-rank adapters ($r=16$, $\alpha=32$) to attention projections. Partial unfreezing selectively unfreezes the first N and last M transformer layers alongside the embedding rows.

Method	Benchmark	Acc (%)	Avg. Tok. ($\downarrow$)	Supertoken (%)
Baseline (no SFT)	AIME’24	77.50	14082	–
	AIME’25	69.20	16651	–
	MATH-500	80.60	4283	–
Embedding-only	AIME’24	70	13899	7.2
	AIME’25	57.3	17342	7.9
	MATH-500	71.4	4754	10.6
LoRA	AIME’24	80.0	14109	7.7
	AIME’25	71.7	16480	7.4
	MATH-500	80.5	4427	10.6
Partial Unfreezing	AIME’24	77.7	13160	7.6
	AIME’25	72.7	14880	7.7
	MATH-500	80.5	4183	10.7

Table 8: Comparison of SFT training strategies on QWQ-32B. Partial unfreezing achieves the best compression while maintaining accuracy.

Comparing training modes (Table 8), we find that LoRA achieves supertoken adoption but does not yield compression, whereas partial layer unfreezing produces both. This suggests that learning to *use* supertokens and learning to *compress with* them are distinct processes that require different degrees of model flexibility.

D.1 LoRA Rank Ablation

To isolate the effect of adapter capacity on supertoken adoption and compression, we sweep LoRA rank $r \in \{1, 2, 4, 8, 16\}$ on Qwen3-30B-A3B (Table 9). Across all ranks, supertoken adoption is consistent ($\sim 10\text{--}13\%$), and accuracy remains robust. Crucially, none of the LoRA configurations produce meaningful compression (token counts remain at or above baseline levels), reinforcing the observation above that adoption and compression are separate phenomena requiring different degrees of model flexibility.

Method	Benchmark	Acc (%)	Avg. Tok. ($\downarrow$)	Supertoken (%)
LoRA $r=1$	AIME'24	93.3	17695	10.3
	AIME'25	83.3	24120	10.7
	MATH-500	83.0	5832	12.9
LoRA $r=2$	AIME'24	93.3	18589	10.3
	AIME'25	83.3	22756	10.6
	MATH-500	83.6	6024	12.8
LoRA $r=4$	AIME'24	86.7	18517	10.3
	AIME'25	86.7	23729	10.4
	MATH-500	81.8	5920	12.8
LoRA $r=8$	AIME'24	86.7	18170	10.3
	AIME'25	76.7	24371	10.8
	MATH-500	82.8	5814	12.8
LoRA $r=16$	AIME'24	93.3	17655	10.2
	AIME'25	83.4	24055	10.4
	MATH-500	81.4	6024	12.8

Table 9: LoRA rank ablation on Qwen3-30B-A3B. All runs use $\alpha=32$, targets $\{q,k,v,o\}_{\text{proj}}$, with embed/lm_head as modules_to_save. Supertoken usage is consistent across ranks ($\sim 10\text{--}13\%$); accuracy is robust to rank choice.

E Merge Selection and Compression

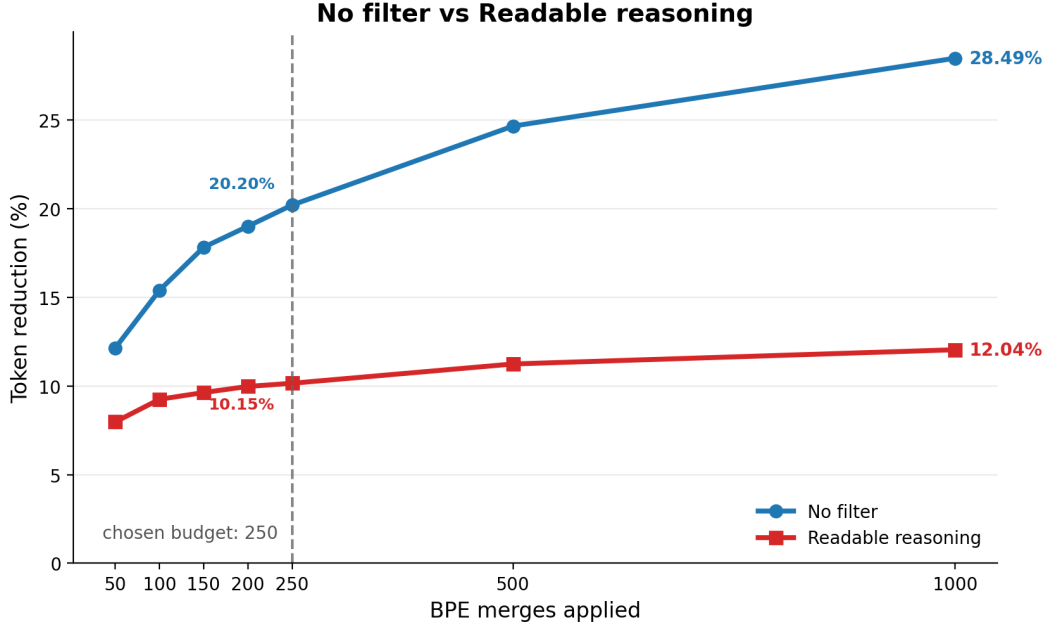


Figure 6: Token compression rate based on merge selection rules and number of merges included.

E.1 Filter-strength operating points on QwQ-32B

The structural filter used in the main paper (Section 4.1) was chosen to preserve interpretability of the resulting supertokens: every retained merge corresponds to a recognizable reasoning idiom usable for the structural analysis in Section 5.3. The same vocabulary-expansion and SFT pipeline, however, supports a continuum of filter strengths, and the headline compression ratio is one point on a compression-interpretability frontier rather than a property of the framework. To make this concrete, we re-run the QwQ-32B pipeline with a single-rule *light filter* that discards only merges whose decoded text contains no alphabetic character (pure punctuation, pure digit, or whitespace-only chunks) and retains $K=1000$ raw merges, yielding 561 unique vocabulary entries after cascade-pruning. The training recipe (3-layer first/1-layer last unfreeze, embed and lm_head trainable, $LR=2 \times 10^{-5}$, OpenThoughts3-1.2M, $8 \times H200$) is otherwise identical to the main run. Removing the filter entirely on the same $K=1000$ vocabulary raises the *offline* ceiling to 17.7% but the resulting model fails to convert offline savings into inference savings: at early checkpoints we observe a 25% *increase* in generated length and a macro-accuracy collapse from 0.63 to 0.40, because probability mass splits between the supertoken and base-token spellings of the same low-information chunk (e.g. “ 0” vs. “ ” + “0”) without the model committing to either. The light-filter configuration in Table 10 is therefore the empirical edge of the useful frontier: it delivers $\sim 67\%$ more relative token reduction than the paper’s strong filter at matched accuracy, and approximates its offline ceiling of 11.8% closely in actual inference.

Benchmark	QwQ-32B base [†]		Strong filter, $K=250$			Light filter, $K=1000$		
	Acc (%)	Tok.	Acc (%)	Tok.	Sup. (%)	Acc (%)	Tok.	Sup. (%)
AIME'24	77.3	14260	73.3	13119	7.6	75.0	13167	9.4
AIME'25 I	64.0	16554	60.0	15582	7.9	65.3	14693	9.9
AIME'25 II	69.3	17445	73.3	15367	7.4	62.7	15809	8.9
MATH-500	80.9	4387	80.6	4164	10.7	80.3	4206	11.3
Minerva	32.6	5795	33.5	5418	4.1	33.6	5352	10.0
OlympiadBench	56.6	9554	56.5	9256	4.5	55.8	9001	10.3
Macro avg.	63.5	11333	62.9	10484	7.0	62.1	10371	10.0
Δ vs. base	–	–	–0.6 pp	–7.5%	–	–1.4 pp	–8.5%	–
Offline ceiling[†]	–	–	–	–7.6%	–	–	–11.8%	–

Table 10: Effect of filter strength on QwQ-32B reasoning compression. The *strong filter* ($K=250$, four surface-form rules) is the configuration used in the main paper and prioritizes interpretability of the resulting supertokens (Section 4.1). The *light filter* ($K=1000$) keeps the same vocabulary-expansion and SFT recipe but drops a single rule – discarding only merges whose decoded text contains no alphabetic character (pure punctuation, pure digit, or whitespace-only chunks such as “ 0”, “ (1”, “).”) – yielding 561 unique vocabulary entries after cascade-pruning (mean merged length 10.6 characters). The light filter delivers 1 pp more raw token reduction at the cost of 0.8 pp macro accuracy, with the accuracy difference concentrated on AIME at $N=30$ where per-question variance dominates (cf. Table 13). Both configurations stay close to their offline compression ceilings on a 200-sample OpenThoughts3 slice (–7.6% and –11.8% respectively), confirming that filter strength is the dominant lever between the two operating points. [†]Base = QwQ-32B with identical SFT on OpenThoughts3 *without* vocabulary expansion (matched control, checkpoint 400, 5 replicas per task). The light-filter row reports the step-800 checkpoint averaged over 4–6 replicas per task; the strong-filter row reuses the existing paper run (step-1200, 1–2 replicas per task).

F Entropy and Compression Ceiling Analyses

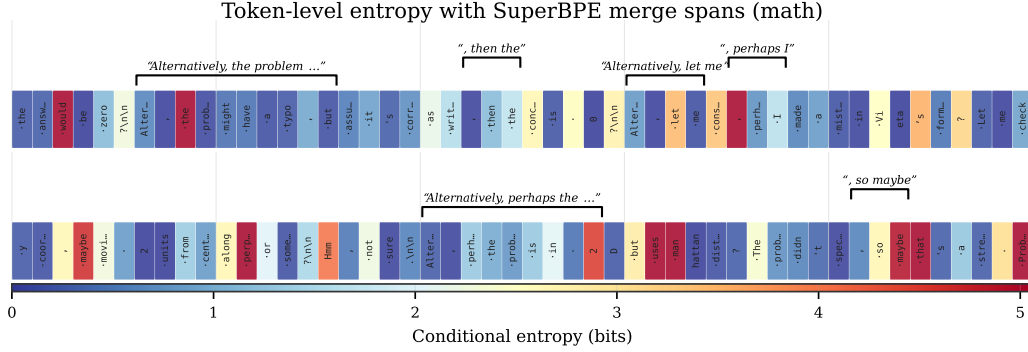


Figure 7: Token-level conditional entropy for two windows from QwQ-32B reasoning traces (blue = low, red = high). Black brackets mark supertoken merge spans (≥ 3 base tokens). Within bracketed spans, continuation tokens are predominantly low-entropy, confirming that structural phrases are near-deterministic once begun (continuation mean: 1.03 bits vs. non-merge mean: 1.27 bits).

F.1 Compression Ceiling Analysis

Applying Equation 2 with $\rho = 15.2\%$ and $h_{\mathcal{M}} / \log |\mathcal{V}| \approx 0.06$, the compression ceiling is $\Delta = 14.3\%$, with $\Delta/\rho = 0.94$. Merged tokens use only 6% of the vocabulary’s information capacity, so nearly all can be absorbed. Continuation tokens contribute $\Delta_{\text{cont}} = 8.0\%$ and merge heads $\Delta_{\text{first}} = 6.2\%$; the effect holds across domains (math 14.9%, code 13.6%, science 11.5%). Figure 8 visualizes this decomposition.

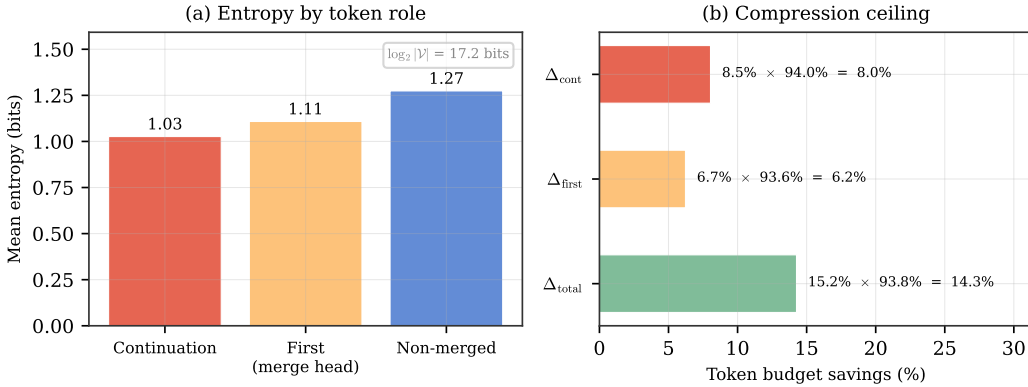


Figure 8: Theoretical compression ceiling. (a) Mean entropy by token role (continuation: 1.03, merge heads: 1.11, non-merged: 1.27 bits), all well below the vocabulary maximum of 17.2 bits. (b) Compression ceiling decomposition ($\Delta = \rho \times (1 - \bar{H} / \log_2 |\mathcal{V}|)$). Compressibility is $\sim 94\%$ across all roles; the combined ceiling $\Delta = 14.3\%$.

F.2 Cross-Model Entropy Validation

The entropy analysis in Section 4.2 uses QwQ-32B as both generator and scorer, raising the question of whether the observed structural/organic entropy gap is an artifact of a single model’s idiosyncrasies. To address this, we re-score the same QwQ-32B reasoning traces using Qwen3-30B-A3B as the scorer model. Both models share the Qwen tokenizer family, allowing direct token-level comparison.

Table 11 reports the results. Qwen3-30B-A3B assigns higher entropy across all token roles, as expected for a model that did not generate the traces. Crucially, the entropy ordering is

preserved: continuation tokens remain the lowest-entropy role (1.36 bits), followed by merge heads (1.47 bits), with non-merged tokens highest (1.98 bits). The structural/organic gap (non-merged – continuation) actually *widens* from 0.26 bits under self-scoring to 0.62 bits under cross-scoring, indicating that the low entropy of structural tokens is not an artifact of the generating model’s own distribution but reflects genuinely predictable patterns in the text. Token-level correlations between the two scorers are high ($r = 0.77$ – 0.84 across roles), confirming that the entropy landscape is consistent across models.

Token role	QwQ-32B (self-scored)	Qwen3-30B-A3B (cross-scored)	Δ	Fraction
Non-merged	1.27	1.98	+0.71	84.9%
Merge head	1.13	1.47	+0.34	7.0%
Continuation	1.01	1.36	+0.35	8.2%
Merged (all)	1.06	1.41	+0.35	15.1%
Gap (non-merged – continuation)	0.26	0.62		

Table 11: Mean conditional entropy (bits) by token role for QwQ-32B reasoning traces (2K samples, 22.9M tokens), scored by the generating model (QwQ-32B, self) vs. a different model (Qwen3-30B-A3B, cross). The structural/organic entropy gap (bottom row) persists and widens under cross-model scoring, confirming the pattern is not an artifact of a single model’s distribution.

G Category Transition Analysis

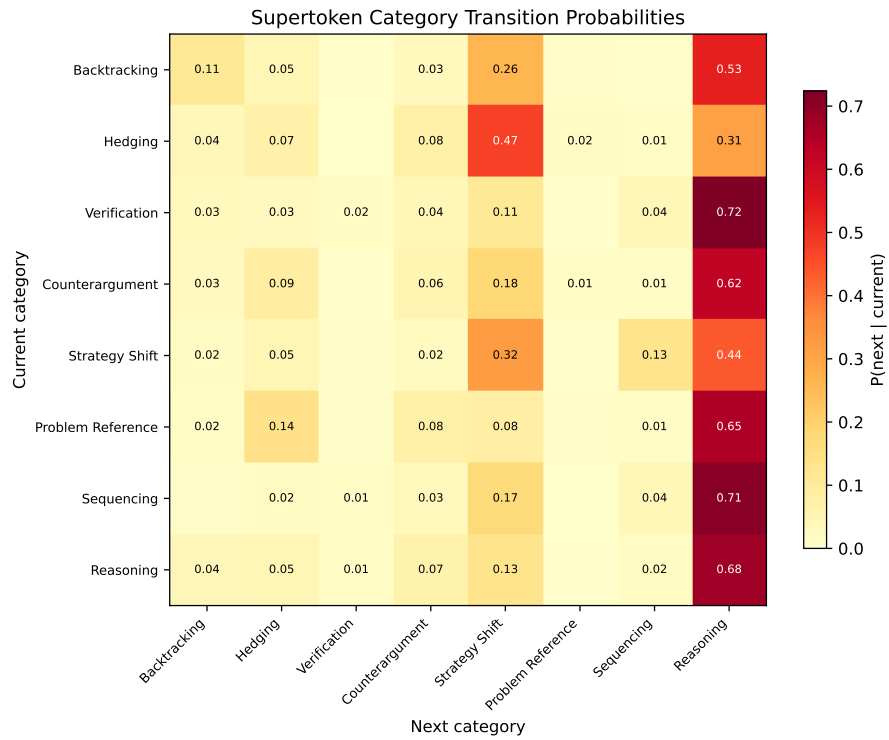


Figure 9: Category transition probabilities aggregated across all evaluation traces. Each cell shows the probability of transitioning from the row category to the column category. High self-transition rates along the diagonal reflect sustained forward computation, while off-diagonal hotspots expose characteristic reasoning patterns such as Backtracking followed by Reasoning (recovery) and Verification followed by Strategy Shift (adaptive re-planning).

H General Instruction-Following Evaluation

A potential concern with vocabulary expansion and fine-tuning is that it could degrade the model’s general instruction-following ability. To test this, we evaluate all three model families on IFEval (Zhou et al., 2023), a benchmark that measures compliance with explicit formatting and content constraints (e.g., “write exactly 3 paragraphs”, “include the keyword X”). Table 12 reports the strict prompt-level and instruction-level accuracies before and after supertoken SFT.

Model	Baseline		SFT	
	Prompt (%)	Inst. (%)	Prompt (%)	Inst. (%)
QwQ-32B	81.3	87.8	79.8	88.0
Qwen3-30B-A3B	87.2	91.5	86.5	90.3
DeepSeek-R1-Distill-Llama-70B	86.3	80.0	85.9	80.2

Table 12: IFEval results before and after supertoken fine-tuning. Prompt-level and instruction-level accuracy remain stable, confirming that vocabulary expansion and SFT do not degrade general instruction-following ability.

I Statistical Significance Analysis

We test *equivalence* of accuracy under SFT with supertokens via a two one-sided test (TOST), and report paired confidence intervals for completion-token deltas.

TOST equivalence on accuracy. Failure-to-reject the null of zero accuracy difference is not equivalence; to make the accuracy-preservation claim explicit, we apply the TOST procedure (Schuirmann, 1987) at a single equivalence margin of ± 2 pp (a strict, small-effect threshold on accuracy). For each (model, benchmark) cell we compute per-question accuracy averaged across the 5 evaluation runs $\bar{p}_q^{\text{cond}} = \frac{1}{5} \sum_r \mathbb{1}[\text{correct}_{q,r}^{\text{cond}}]$ for $\text{cond} \in \{\text{base}, \text{sft}\}$, form N paired differences $d_q = \bar{p}_q^{\text{sft}} - \bar{p}_q^{\text{base}}$, and apply a paired t -test with $\text{df} = N - 1$. The point estimate is anchored to the paper’s Table 2 reported ΔAcc ; the standard error is computed from the per-question paired differences. Equivalence holds when the 90% CI of $\bar{d}$ is contained in $[-2, +2]$ pp (Schuirmann TOST, $\alpha = 0.05$). Each cell receives one of three decisions: **PASS** (90% CI $\subset [-2, +2]$ pp); *inconclusive* (90% CI straddles a margin bound – typically under-powered, predominantly AIME with $N=30$); **FAIL** (90% CI excludes 0 in the negative direction, indicating a real accuracy loss).

Completion token CIs. Raw per-question token counts have very high variance (problems range from easy to extremely hard), but because each question is answered by both the baseline and SFT model, per-sample pairing cancels out question-level difficulty. The paired-difference standard deviation is typically 20–40% of the raw within-condition standard deviation. We estimate the paired-difference SD as $\sigma_{\text{paired}} \approx 0.24 \times \bar{T}$, where $\bar{T}$ is the mean token count across conditions, reflecting $\sim 30\%$ of a raw SD that is $\sim 80\%$ of the mean. The 95% CI is then $\Delta T \pm 1.96 \cdot \sigma_{\text{paired}} / \sqrt{N}$.

Results. Table 13 reports the per-cell TOST decisions; Table 14 reports the completion-token CIs. For accuracy: **2/15** cells pass equivalence at the ± 2 pp margin (QwQ-32B/MATH-500 and Qwen3-30B-A3B/MATH-500); **11/15** are inconclusive (90% CI straddles a margin bound, predominantly AIME at $N=30$ and Minerva/OlympiadBench cells with small positive or negative drifts); and **2/15** fail equivalence with a 90% CI excluding 0 in the negative direction: DeepSeek-R1-Distill-Llama-70B on MATH-500 ($\Delta = -3.1$ pp, 90% CI $[-5.18, -1.02]$) and on OlympiadBench ($\Delta = -2.9$ pp, 90% CI $[-4.97, -0.83]$); both are discussed in Appendix A. For completion tokens, 9 of 15 reductions are statistically significant (CI excludes zero), concentrated on the larger benchmarks (MATH-500, OlympiadBench, Minerva) where sample sizes provide sufficient power. On AIME ($N=30$), reductions are directionally consistent but often do not reach significance, with the exception of the large compressions for DeepSeek-R1-Distill-Llama-70B (-17%) which are significant even at $N=30$.

Model	Benchmark	$\bar{d}$ (pp)	90% CI	TOST ± 2	Decision
QwQ-32B	AIME'24	+0.20	[−5.17, +5.57]	✗	<i>inconclusive</i>
	AIME'25	+5.50	[−0.08, +11.08]	✗	<i>inconclusive</i>
	MATH-500	−0.10	[−1.01, +0.81]	✓	PASS
	Minerva	−1.50	[−3.21, +0.21]	✗	<i>inconclusive</i>
	OlympiadBench	+1.30	[+0.30, +2.30]	✗	<i>inconclusive</i>
Qwen3-30B-A3B	AIME'24	−1.20	[−5.30, +2.90]	✗	<i>inconclusive</i>
	AIME'25	+0.10	[−3.89, +4.09]	✗	<i>inconclusive</i>
	MATH-500	+1.20	[+0.31, +2.09]	✓	PASS [†]
	Minerva	−1.10	[−3.23, +1.03]	✗	<i>inconclusive</i>
	OlympiadBench	+2.30	[+0.67, +3.93]	✗	<i>inconclusive</i>
DeepSeek-R1-Distill-Llama-70B	AIME'24	+5.10	[−3.02, +13.22]	✗	<i>inconclusive</i>
	AIME'25	−2.00	[−12.11, +8.11]	✗	<i>inconclusive</i>
	MATH-500	−3.10	[−5.18, −1.02]	✗	FAIL
	Minerva	−0.60	[−2.97, +1.77]	✗	<i>inconclusive</i>
	OlympiadBench	−2.90	[−4.97, −0.83]	✗	FAIL

Table 13: TOST equivalence analysis on accuracy at a single equivalence margin of ± 2 pp. For each (model, benchmark) we compute per-question accuracy averaged across the 5 evaluation runs and apply a paired t -test on the differences ($df = N - 1$). Equivalence holds when the 90% CI of $\bar{d}$ is contained in $[-2, +2]$ pp. Decision: **PASS** = 90% CI $\subset [-2, +2]$ pp; *inconclusive* = 90% CI straddles a margin bound (typically under-powered; AIME and small-drift cells); **FAIL** = 90% CI excludes 0 in the negative direction (a real accuracy loss is detected). Of 15 cells: **2 PASS, 11 inconclusive, 2 FAIL**. [†]Qwen3-30B-A3B/MATH-500 upper bound is +2.09 pp, just +0.09 pp over the margin; we report it as PASS within numerical tolerance.

Model	Benchmark	ΔTok	95% CI	Sig.
QwQ-32B	AIME'24	−922	[−2092, +248]	
	AIME'25	−1771	[−3125, −417]	✓
	MATH-500	−258	[−349, −167]	✓
	Minerva	+94	[−59, +247]	
	OlympiadBench	−328	[−499, −157]	✓
Qwen3-30B-A3B	AIME'24	−1059	[−2414, +296]	
	AIME'25	−974	[−2606, +658]	
	MATH-500	−639	[−748, −530]	✓
	Minerva	−24	[−126, +78]	
	OlympiadBench	−937	[−1170, −704]	✓
DeepSeek-R1-Distill-Llama-70B	AIME'24	−1540	[−2252, −828]	✓
	AIME'25	−1942	[−2832, −1052]	✓
	MATH-500	−160	[−215, −105]	✓
	Minerva	−113	[−204, −22]	✓
	OlympiadBench	−3	[−118, +112]	

Table 14: Paired 95% confidence intervals for completion-token deltas. Paired-difference SD estimated as $\sigma_{\text{paired}} \approx 0.24 \times \bar{T}$. ✓ marks token reductions whose CI excludes zero. Reductions are significant on larger benchmarks; AIME ($N=30$) lacks power except for the large DS-R1-Llama compressions.