S^3: SPECTRAL NULL-SPACE SWAP MAKES REASONING MODELS EFFICIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs trained with Chain-of-thought excel in reasoning capability, but often come with excessive token cost. We find that the core of reasoning capacity lies in the Thinking model's weight component within the null space of a projection defined by the corresponding Non-thinking model's dominant singular directions, and removing the subspace component can largely improve reasoning efficiency without hurting the accuracy gained during thinking-mode post-training. Unlike existing efforts that mostly operate within the dominant subspace, we are the first to unveil the critical role of the null space and harness it for model optimization. Motivated by this finding, we propose **Spectral Null-Space Swap** (S^3), a training-free composition of paired Non-thinking and Thinking checkpoints. Our method keeps the Non-thinking model inside its own dominant subspace and takes the Thinking checkpoint outside it, improving reasoning efficiency while maintaining accuracy. We extensively evaluate S^3 on 2B–30B dense and mixture-of-experts (MoE) architectures spanning 28 evaluation environments across **mathematical, multimodal, and audio reasoning domains**. S^3 establishes new empirical Pareto Frontiers among training-free model composition strategies: across all settings, it reduces inference token overhead by an average of 27.4% compared to full Thinking models while simultaneously improving overall task accuracy by 1.0 percentage point (e.g., yielding +8.3% accuracy on HMMT25 alongside a 33.0% token speedup). We further use attention entropy for explanation and find that the retained component produces more concentrated attention, and we use a simplified analytical model about optimization to demonstrate why null-space can effectively reduce attention entropy, thereby improving the efficiency of reasoning.

1 INTRODUCTION

Scaling language models enabled few-shot adaptation Brown et al. (2020), instruction tuning aligned them with user intent Ouyang et al. (2022), and chain-of-thought (CoT) prompting made reasoning explicit Wei et al. (2022). Recent models extend deliberate reasoning across text, vision, and audio Yang et al. (2025); Bai et al. (2025); Xu et al. (2025), while agents interleave reasoning with actions and observations Yao et al. (2023); Liu et al. (2024); Xu et al. (2026b); Zhang et al. (2026c); Guo et al. (2026a;b) to handle extended interactions, multi-agent coordination, and sequential decision-making Huang et al. (2026a); Ma et al. (2026b; 2025); Zhu et al. (2025; 2026); Ma et al. (2026a). However, overly lengthy CoT imposes heavy computational costs as workloads scale, making *efficient reasoning* essential.

Post-training—especially reinforcement learning in thinking mode—improves reasoning through multi-domain fine-tuning, policy optimization, and reward modeling Liang et al. (2025b); Liu et al. (2026); Xu et al. (2026c); Wang et al. (2026); Pan et al. (2026); Snell et al. (2025), but inevitably inflates token consumption. Prior work explores explicit budgets Muennighoff et al. (2025), adaptive stopping, token pruning, and weight interpolation to balance accuracy and efficiency. Nevertheless, a core question remains: *where in weight space does a Thinking model differ functionally from its Non-thinking counterpart, and which part of that difference can be removed without incurring full generation costs?*

In spectral coordinates of the Non-thinking model, let W_0 and W_t be paired Non-thinking and Thinking weights, with post-training difference $\Delta W = W_t - W_0$. We decompose ΔW into com-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

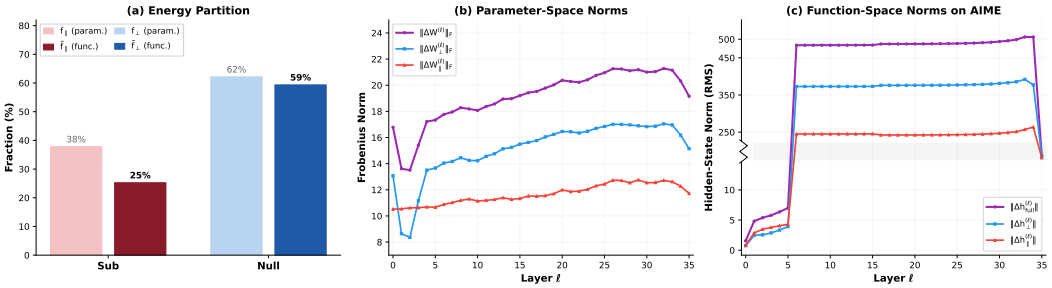


Figure 1: **This figure compares the parallel and perpendicular components of the thinking difference in parameter and function spaces.** All panels use the Qwen3-4B-Instruct and Qwen3-4B-Thinking. Parameter space denotes weight differences; function space denotes hidden-state offsets relative to Qwen3-4B-Instruct on AIME24/25. (a) Global average results. (b) Per-layer parameter-space results. (c) Per-layer function-space results on the same inputs.

ponents aligned with and complementary to W_0's singular subspace, comparing their Frobenius energies with induced hidden-state changes. Figure 1 reveals a structural asymmetry for Qwen3-4B on AIME24/25: the aligned component holds sizable parameter energy but induces weak functional shifts, whereas the complementary component drives most functional change despite lower energy. Thus, weight changes along directions the Non-thinking model already occupies have little functional effect; function resides mostly in the complement. Unlike prior studies that primarily manipulate components within the dominant subspace, this finding underscores the untapped potential of the complementary space and motivates *keeping Non-thinking in the former and importing Thinking in the latter*.

We instantiate this idea as **Spectral Null-Space Swap** (S^3). For each target matrix, the leading singular directions of W_0 define a protected subspace $\mathcal{S}_\rho$; the composed weight is $W_\rho = P_{\mathcal{S}_\rho}(W_0) + (I - P_{\mathcal{S}_\rho})(W_t)$. The protected region therefore keeps the Non-thinking model's structure, while its complement is taken from the Thinking model. Existing weight-space approaches average or interpolate checkpoints, resolve merging conflicts, or retain Base-aligned updates. S^3 instead makes an asymmetric, subspace-specific choice about *which model supplies each region* without the additional training.

We evaluate paired 2B–30B dense and MoE models across text, vision–language, and audio tasks. Across diverse architectures and multimodal benchmarks, S^3 achieves substantial token efficiency while maintaining or improving accuracy relative to Thinking baselines. Crucially, several configurations establish empirical Pareto-optimal accuracy–token trade-offs among evaluated merging methods. For instance, Qwen3-4B-S^3 gains 8.3 percentage points in accuracy on HMMT25 alongside a 33.0% token reduction; on MathVista, Qwen3-VL-4B-S^3 reaches 78.05% accuracy with 31.9% fewer tokens. The $\rho = 1$ ablation isolates spectral components: its null-space model nearly matches Thinking accuracy with over a quarter fewer tokens, whereas the aligned-subspace model degrades toward Non-thinking.

Finally, we use attention entropy to examine why this decomposition changes reasoning behavior. Across evaluated datasets, the null-space model concentrates attention more than Non-thinking models, whereas adding only the aligned component diffuses attention, yielding the empirical ordering $H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}$. Our simplified analytical model provides an explanation of when this ordering arises, offering a conditional prediction for small perturbation that matches observed attention patterns and shorter reasoning traces.

Our main contributions can be summarized as follows:

- **Spectral finding.** We characterize the post-training thinking difference in both parameter and function spaces and identify a mismatch between their energy shares, showing that components along directions already occupied by Non-thinking models yield less functional impact, which instead largely resides in the complementary space.
- **Selective composition.** We introduce **Spectral Null-Space Swap** (S^3), a training-free composition of Non-thinking and Thinking checkpoints. It preserves Non-thinking's lead-

ing spectral directions and imports Thinking's complementary component. This transfer differs from whole-checkpoint interpolation and methods retaining base-aligned updates.

- **Accuracy–efficiency results.** We evaluate five paired model families spanning 2B–30B dense and MoE architectures across text, vision–language, and audio reasoning. Across settings, S^3 achieves multiple Pareto-Optimal operating points and uses 27.4% fewer tokens on average than Thinking while improving mean accuracy by 1.0 percentage point.
- **Mechanistic explanation.** We introduce attention entropy to explain how S^3 alters model behaviors via spectral decomposition. Specifically, we observe the persistent ordering $H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}$ across evaluated tasks and provide a simplified analytical model.

2 Related Work

2.1 Efficient Reasoning and Chain-of-Thought Compression

Chain-of-thought (CoT) improves LLM reasoning but increases decoding cost. Existing approaches address this trade-off through adaptive computation and post-training, including adaptive stopping, token pruning, partial verification, process-reward learning, and reinforcement-learning-based optimization Huang et al. (2026b); Li et al. (2026b); Dong et al. (2026); Fan et al. (2026; 2025); Yang et al. (2026); Liang et al. (2025a); Cai et al. (2026); Cai & Sugiyama (2026). In contrast, our method operates directly in weight space: it preserves the Non-thinking model in a protected spectral subspace and transfers only the complementary component from its corresponding Thinking model. This retains the Thinking model's reasoning capability while reducing CoT overhead, without additional training or decoding changes.

2.2 Weight-Space Model Composition and Spectral Subspace Transfer

Weight-space merging combines fine-tuned models without retraining, evolving from checkpoint averaging and task arithmetic to interference reduction via sign resolution and delta sparsification Wortsman et al. (2022); Ilharco et al. (2023); Yadav et al. (2023); Yu et al. (2024). For efficient reasoning, direct Non-thinking–Thinking interpolation offers an accuracy–cost trade-off without extra training Wu et al. (2026a). Spectral methods exploit parameter subspaces, including task-matrix analysis, LoRA-SVD alignment, shared/task-specific subspace separation, and base-aligned RL update decomposition Gargiulo et al. (2025); Stoica et al. (2025); Marczak et al. (2025); Zhang et al. (2026b). Our approach instead preserves the Non-thinking model's leading spectral subspace while transferring only the complementary component from the Thinking model, using Non-thinking as an anchor and Thinking as a capability donor for selective reasoning transfer rather than multi-task fusion.

2.3 Attention Entropy and Reasoning Dynamics

Attention entropy characterizes attention concentration and has been linked to Transformer attention dynamics and geometry Ghader & Monz (2017); Zhai et al. (2023); Ryu & Lewis (2021). Recent work further relates attention entropy to reasoning dynamics and chain-of-thought trajectories Wen et al. (2025); Zhang et al. (2026a); Guan et al. (2026); Shang et al. (2026). We observe consistent reductions in attention entropy across models and datasets, consistent with prior studies linking more concentrated attention patterns to effective contextual processing and reasoning, as well as recent work relating entropy dynamics to reasoning efficiency Wen et al. (2025); Zhang et al. (2025); Xiong et al. (2026). Beyond this empirical observation, we provide the first theoretical analysis connecting null-space projection in parameter space to attention entropy.

3 Method: Spectral Null-Space Swap

This section develops the method in three steps. Section 3.1 asks where a Thinking model differs from its Non-thinking counterpart in weight space, finding that functional differences lie mostly outside the Non-thinking model's occupied directions. Section 3.2 converts this finding into an operator that preserves the Non-thinking model's leading spectral directions while importing the remainder from the Thinking model, controlled by a single parameter. Section 3.3 examines full protection,

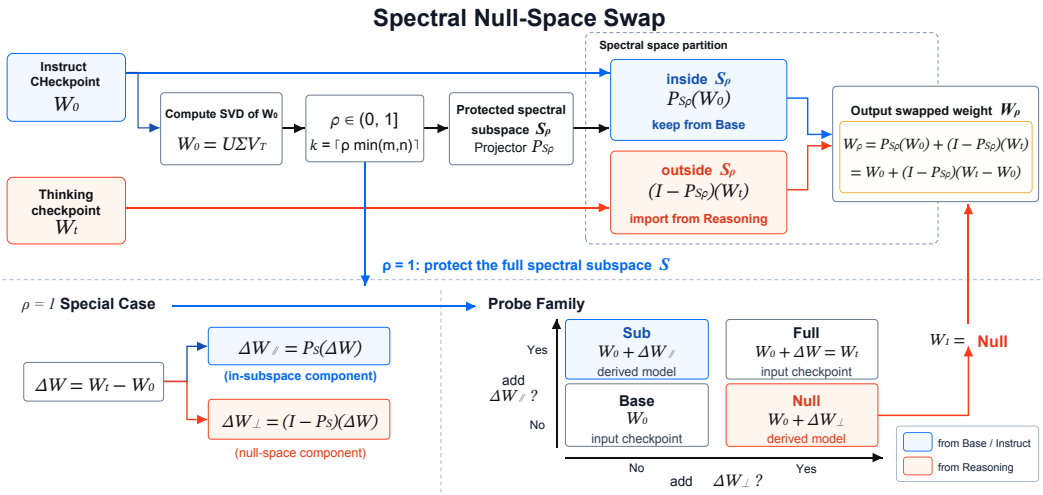


Figure 2: **Overview of our method: Spectral Null-Space Swap and the $\rho = 1$ probe family.** The swapped weight preserves the protected spectral component of the Non-thinking model and imports the complementary component of the Thinking model. The four probes isolate the contributions of $\Delta W_\parallel$ and $\Delta W_\perp$.

where the operator transfers only expressive complements and forms a four-variant family differing solely in retained components—laying the groundwork for the interpretability analysis in Section 5.

3.1 Insight: Where the Thinking Difference Lives

Let W_0 and W_t denote the weights of a paired Non-thinking and Thinking checkpoints from the same model family. The two checkpoints always descend from the same pretrained model and differ in whether they produce an extended chain of thought before answering Yang et al. (2025); Bai et al. (2025). All the thinking adds is the difference $\Delta W = W_t - W_0$, which is available from the two released checkpoints alone.

We measure ΔW in two ways. Its *parameter-space* size is its Frobenius energy; its *function-space* size is the change it induces in the network's computation. To separate the two, let $W_0 = U\Sigma V^\top$ be a compact SVD and $P_S(X) = U(U^\top XV)V^\top$ the orthogonal projection onto the subspace S spanned by the singular directions of W_0. Since $P_S(W_0) = W_0$, this splits the difference into the part that lies where the anchor weight already has support and its complement—for the rectangular attention projections, exactly the null-space component:

$$\Delta W = \underbrace{P_S(\Delta W)}_{\Delta W_\parallel} + \underbrace{(I - P_S)(\Delta W)}_{\Delta W_\perp}, \qquad \|\Delta W\|_F^2 = \|\Delta W_\parallel\|_F^2 + \|\Delta W_\perp\|_F^2, \tag{1}$$

with parameter-space share $f_\parallel^{(\ell)} = \|\Delta W_\parallel^{(\ell)}\|_F^2 / \|\Delta W^{(\ell)}\|_F^2$ and $f_\perp^{(\ell)} = 1 - f_\parallel^{(\ell)}$.

For the function-space counterpart, write $h_M^{(\ell)}(x) \in \mathbb{R}^{T_x \times d}$ for the hidden states of model M on input x at layer ℓ, where $\ell = 0$ is the embedding output and $\ell = 1, \ldots, L$ are the outputs of the transformer layers, and let

$$\Delta h_*^{(\ell)}(x) = h_{M_*}^{(\ell)}(x) - h_{W_0}^{(\ell)}(x), \qquad M_* \in \{W_0 + \Delta W_\parallel,\ W_0 + \Delta W_\perp,\ W_0 + \Delta W\} \tag{2}$$

be the shift that each variant induces relative to the anchor. Because the shifts are read off actual forward passes, they capture the full nonlinear effect of a weight change rather than a first-order Jacobian approximation. Averaging the squared token-wise shifts over the $N = \sum_x T_x$ tokens of a shared input set X gives the functional norm and the corresponding share

$$\left\|\Delta h_*^{(\ell)}\right\|^2 = \frac{1}{N} \sum_{x \in X} \sum_{t=1}^{T_x} \left\|\Delta h_*^{(\ell)}(x)_t\right\|_2^2, \qquad \tilde{f}_*^{(\ell)} = \frac{\|\Delta h_*^{(\ell)}\|^2}{\|\Delta h_{\text{full}}^{(\ell)}\|^2}. \tag{3}$$

Figure 1 compares the two decompositions. The split is stable across the middle and late layers, while the parameter-space and function-space measures consistently show that the component inside S accounts for much of the Frobenius energy but far less of the functional change, while the complementary component has similar shares under both measures. This suggests that a substantial portion of the difference lies along directions the anchor already occupies, yet contributes little to the model's actual computation. Thus, the functionally critical part of the thinking difference lies mainly outside the Non-thinking model's spectral subspace, and the in-subspace component is largely redundant. This identifies which side the composed model should keep from the anchor and lays the foundation for model composition. Section 3.2 turns this observation into a concrete operator.

3.2 The Spectral Null-Space Swap

Section 3.1 splits the difference along the singular directions of W_0. We now turn that split into an operator by letting only the leading fraction of those directions play the role of $\mathcal{S}$. For $\rho \in (0, 1]$ let U_k, V_k collect the leading $k = \lceil \rho \min(m, n) \rceil$ singular vectors and

$$P_{\mathcal{S}_\rho}(X) = U_k(U_k^\top X V_k)V_k^\top, \qquad \mathcal{S}_\rho = \mathrm{span}(U_k) \otimes \mathrm{span}(V_k) \subseteq \mathcal{S}. \tag{4}$$

As Figure 2 illustrates, the **Spectral Null-Space Swap** keeps the Non-thinking model inside $\mathcal{S}_\rho$ and takes the Thinking model outside it,

$$W_\rho \;=\; P_{\mathcal{S}_\rho}(W_0) \;+\; (I - P_{\mathcal{S}_\rho})(W_t). \tag{5}$$

Equation equation 5 admits two equivalent interpretations. From a subspace perspective,

$$P_{\mathcal{S}_\rho}(W_\rho) = P_{\mathcal{S}_\rho}(W_0), \qquad (I - P_{\mathcal{S}_\rho})(W_\rho) = (I - P_{\mathcal{S}_\rho})(W_t), \tag{6}$$

which shows that the two spectral regions are inherited from different sources: inside the protected subspace $\mathcal{S}_\rho$, the composed weight exactly matches the Non-thinking model, whereas on the orthogonal complement it exactly matches the Thinking model. Thus, $\mathcal{S}_\rho$ acts as the interface separating the structure preserved from W_0 from the component imported from W_t.

The same construction can be viewed from the perspective of transferring the difference. Rewriting equation 5 gives

$$W_\rho = W_0 + (I - P_{\mathcal{S}_\rho})(W_t - W_0), \tag{7}$$

so the composed model can be obtained by starting from W_0, decomposing the thinking difference $\Delta W = W_t - W_0$ into components inside and outside $\mathcal{S}_\rho$, discarding the former, and transferring only the latter. In this view, the operation does not interpolate the two checkpoints globally; instead, it selectively injects the part of the thinking difference that lies outside the protected spectral subspace of the Non-thinking model.

3.3 Specialization at $\rho = 1$: The Probe Family

Setting $\rho = 1$ in equation 7 removes the only free parameter: $k = \min(m, n)$, hence $U_k = U$ and $V_k = V$, and the swap returns

$$W_1 \;=\; W_0 + (I - P_{\mathcal{S}})(W_t - W_0) \;=\; W_0 + \Delta W_\perp, \tag{8}$$

a transfer of exactly the component of ΔW that W_0 cannot represent (for the $m > n$ MLP matrices, the left-null-space component instead, so the expression is well defined for every target weight). Keeping or discarding each of the two components of ΔW then gives four variants,

$$\underbrace{W_0}_{\text{Base}}, \qquad \underbrace{W_0 + \Delta W_\parallel}_{\text{Sub}}, \qquad \underbrace{W_0 + \Delta W_\perp}_{\text{Null}}, \qquad \underbrace{W_0 + \Delta W}_{\text{Full}}, \tag{9}$$

of which *Base* and *Full* are the released checkpoints and *Sub* and *Null* are derived from them. By equation 8 the **Spectral Null-Space Swap** is the Null entry, so the whole family is available from the two checkpoints alone, with no training, data, or roll-outs.

The construction is also convenient for interpretability: the four variants of equation 9 differ only in which spectral component of a single difference is applied, and the decomposition is exact and local to each weight matrix, so an observed effect can be attributed to specific directions in a specific layer. Section 5 uses the family in this way.

4 EXPERIMENTS

In this section, we first describe the evaluation settings and baselines across text, vision–language, and audio reasoning tasks in Section 4.1. We then evaluate the cross-modal accuracy–efficiency trade-off of our method, study the effect of the protected-subspace ratio ρ, and isolate the respective contributions of the subspace and null-space components in Section 4.2, Section 4.3 and Section 4.4. Additional selected-result visualizations and analyses, including Pareto-frontier plots and extended ratio sweeps, are provided in Appendix C.

Table 1: **Main results across text, vision–language, and audio reasoning benchmarks.** We report accuracy (%) and average generated tokens per problem. **Bold** indicates the best value in each column among TIES, MI-0.8, and S^3. In each Δ row, accuracy is reported as an absolute change and token usage as a relative change with respect to the corresponding Thinking model.

Large Language Model	**AIME25**		**HMMT25**		**CMIMC25**		**Olympiad-Bench**	
	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓
Qwen3-4B								
Qwen3-4B-Instruct	45.8	6,785	35.0	8,621	32.5	8,461	77.3	4,468
Qwen3-4B-Thinking	71.7	20,704	46.7	24,636	58.1	24,415	83.8	14,086
Qwen3-4B-TIES	**73.3**	16,377	50.0	**16,049**	54.4	21,748	**85.8**	10,238
Qwen3-4B-MI-0.8	71.7	17,270	**58.3**	19,729	**56.3**	22,902	85.5	11,460
Qwen3-4B-S^3-0.8	**73.3**	**15,004**	55.0	16,496	55.0	**21,266**	83.8	**9,745**
Δ *vs.* Thinking	+1.6	−27.5%	+8.3	−33.0%	−3.1	−12.9%	+0.03	−30.8%
Qwen3-30B-A3B								
Qwen3-30B-A3B-Instruct	60.0	6,851	45.8	7,558	45.0	7,478	82.2	4,305
Qwen3-30B-A3B-Thinking	84.2	17,579	68.3	21,093	64.4	21,979	85.8	12,359
Qwen3-30B-A3B-TIES	72.5	**10,472**	52.5	**15,555**	58.1	16,068	84.0	**8,560**
Qwen3-30B-A3B-MI-0.8	76.7	15,648	**65.0**	19,432	**68.1**	19,632	**86.9**	10,899
Qwen3-30B-A3B-S^3-0.8	**80.8**	12,536	62.5	16,424	66.3	**15,988**	88.4	9,238
Δ *vs.* Thinking	−3.37	−28.7%	−5.8	−22.1%	+1.85	−27.3%	+2.55	−25.3%
Vision Language Model	**GSM8K**		**MMLU**		**MMMU**		**MathVista-testmini**	
Qwen3-VL-2B								
Qwen3-VL-2B-Instruct	87.88	918.5	65.68	484.0	46.00	6,665.8	57.73	2,235.4
Qwen3-VL-2B-Thinking	92.75	2,295.4	80.15	3,228.0	55.50	8,075.4	66.90	3,623.8
Qwen3-VL-2B-TIES	**93.00**	1,877.9	80.59	1,962.4	**60.00**	7,601.8	68.23	3,085.6
Qwen3-VL-2B-MI-0.8	92.88	1,972.4	80.59	2,333.7	55.50	7,836.5	67.58	3,349.8
Qwen3-VL-2B-S^3-0.8	91.62	**1,550.1**	**80.92**	**1,754.8**	57.50	**7,513.0**	**68.58**	**2,784.2**
Δ *vs.* Thinking	−1.13	−32.5%	+0.77	−45.6%	+2.00	−7.0%	+1.68	−23.2%
Qwen3-VL-4B								
Qwen3-VL-4B-Instruct	96.37	266.1	78.62	57.8	50.25	597.4	73.35	1,104.9
Qwen3-VL-4B-Thinking	96.00	1,403.0	90.02	1,823.9	80.00	5,152.7	76.02	2,843.9
Qwen3-VL-4B-TIES	**96.50**	1,258.9	88.49	1,284.8	80.75	4,521.2	**79.90**	2,273.9
Qwen3-VL-4B-MI-0.8	96.37	1,243.4	**89.47**	1,268.3	78.75	4,375.1	77.35	2,181.4
Qwen3-VL-4B-S^3-0.8	96.00	**1,139.6**	88.49	**1,167.3**	**82.50**	**4,094.8**	78.05	**1,936.1**
Δ *vs.* Thinking	0.00	−18.8%	−1.53	−36.0%	+2.50	−20.5%	+2.03	−31.9%
Omni Language Model	**GSM8K**		**MMLU**		**MMAR**		**MMSU**	
Qwen3-Omni-30B-A3B								
Qwen3-Omni-30B-A3B-Instruct	98.00	312	90.40	105	35.20	5,652	66.75	1,022
Qwen3-Omni-30B-A3B-Thinking	97.60	936	92.80	1,166	63.30	982	70.33	622
Qwen3-Omni-30B-A3B-TIES	97.20	860	92.90	693	64.97	799	66.47	530
Qwen3-Omni-30B-A3B-MI-0.8	**97.40**	845	**93.20**	1,036	**66.33**	820	67.70	643
Qwen3-Omni-30B-A3B-S^3-0.8	96.88	**754.2**	91.89	**608.4**	64.70	**740**	**68.18**	**526**
Δ *vs.* Thinking	−0.72	−19.4%	−0.91	−47.8%	+1.40	−24.6%	−2.15	−15.4%

4.1 EXPERIMENTAL SETUP

Settings. We evaluate **Spectral Null-Space Swap** on paired Non-thinking and Thinking checkpoints from the Qwen3 family Yang et al. (2025). Our evaluation covers text, vision–language, and audio reasoning tasks. Specifically, we use Qwen3-4B, Qwen3-30B-A3B, Qwen3-VL-2B, Qwen3-VL-4B, Qwen3-Omni-30B-A3B as our backbone models Bai et al. (2025); Xu et al. (2025). This

Table 2: Effect of the protected-subspace ratio in Qwen3-4B-S^3 on the four benchmarks reported in the main table. We report accuracy (%) and average generated tokens per problem. **Bold** indicates the best value in each column.

Model	AIME25		HMMT25		CMIMC25		Olympiad-Bench	
Qwen3-4B-S^3-1.0	69.2	15,129	45.8	**15,920**	50.6	**17,863**	83.0	**9,553**
Qwen3-4B-S^3-0.9	71.7	15,843	49.2	16,733	53.1	20,567	83.5	9,607
Qwen3-4B-S^3-0.8	73.3	**15,004**	**55.0**	16,496	**55.0**	21,266	83.8	9,745
Qwen3-4B-S^3-0.7	75.8	18,082	45.0	22,377	48.1	22,186	**84.0**	10,265
Qwen3-4B-S^3-0.6	78.3	18,542	49.2	22,481	50.6	22,798	83.8	11,607
Qwen3-4B-S^3-0.5	**79.2**	19,264	54.2	23,361	51.3	23,743	83.6	12,035

design evaluates both cross-modal generalization and compatibility with dense and MoE architectures. We report **pass@1** for single-sample correctness and **avg@4** for average performance over four generations. Additional model configurations are provided in Appendix B.1.

Baselines. We compare our method with the original Non-thinking and Thinking checkpoints from each paired model family. We further compare against MI-0.8, a direct Instruct–Thinking weight-interpolation baseline using the 0.8 interpolation setting Wu et al. (2026a), and TIES-Merging, which resolves parameter conflicts through trimming and sign election Yadav et al. (2023). All methods operate on the same paired checkpoints and are evaluated using identical prompts and decoding configurations.

4.2 Main Results: Cross-Modal Accuracy–Efficiency Trade-offs

Table 1 shows that our method consistently provides a favorable accuracy–efficiency trade-off across text, vision–language, and audio reasoning. Compared with Non-thinking models, it substantially improves reasoning while maintaining markedly shorter outputs than Thinking models. On text tasks, our method recovers most Thinking gains—and exceeds them on several benchmarks—while shortening chains of thought. Across vision–language and audio domains, it frequently achieves the lowest token usage among merging baselines while preserving competitive accuracy. Thus, **Spectral Null-Space Swap** transfers reasoning capability without the full decoding overhead. Compared with TIES and MI-0.8, our method exhibits a superior Pareto frontier; additional benchmarks and detailed trade-off analyses appear in Appendix B.2.

4.3 Ablation on the Protected-Subspace Ratio ρ

We study the effect of the protected-subspace ratio ρ, which controls the fraction of leading singular directions of the Non-thinking checkpoint retained during spectral subspace swap. Table 2 reports the results of Qwen3-4B-S^3 across $\rho \in [0.5, 1.0]$.

Overall, decreasing ρ transfers a larger component from the Thinking model and generally increases reasoning capability, but also leads to longer generations. This trend is particularly clear on AIME25, where smaller ratios improve accuracy while increasing token usage. In contrast, larger ratios yield the shortest outputs on several benchmarks, reflecting stronger preservation of the efficient Non-thinking model behavior. We use $\rho = 0.8$ as the default configuration because it offers a balanced trade-off, achieving strong reasoning performance while avoiding the substantially longer trajectories observed at smaller ratios. Additional ablation results are shown in Appendix B.3.

4.4 Ablation on Subspace and Null-Space Components

Although the method in Section 3 performs well across benchmarks, we must rule out that the subspace component alone would suffice. We therefore compare the *Sub* and *Null* models against the *Base* and *Full* models on math benchmarks. Table 3 shows that the *Sub* model stays close to the *Base* model, while the *Null* model matches the *Full* model. Therefore, the reasoning gain is concentrated in the null-space component.

Table 3: **Accuracy and token length across five math benchmarks generated by the probe family.** The *Sub* model is almost identical to the *Base* model. The *Null* model matches the *Full* model's average accuracy while generating fewer tokens.

Model	AIME'24	AIME'25	AMC'23	CMIMC'25	HMMT'25	Mean Acc	Mean len
Base	61.7	50.4	94.7	30.6	33.3	54.1	8,280
Sub	69.2	56.2	95.9	36.2	35.4	58.6	8,891
Null	77.1	74.6	99.1	51.2	49.2	70.2	15,071
Full	75.4	70.0	99.1	53.4	53.3	70.3	20,269

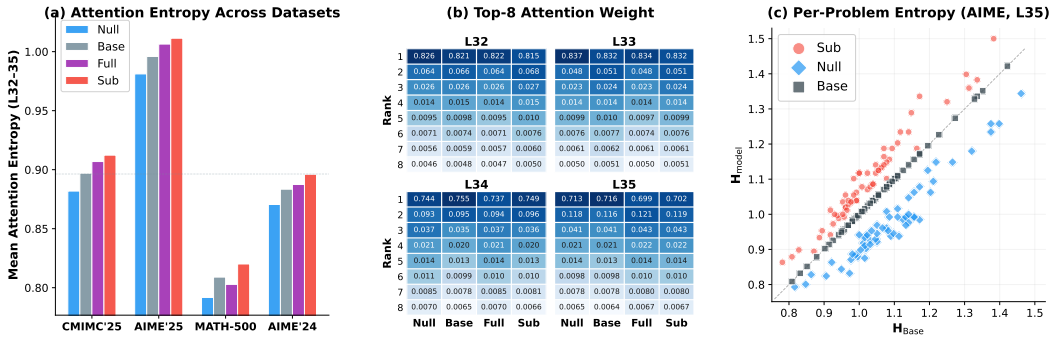


Figure 3: **Attention entropy analysis across four model variants.** (a) Mean attention entropy averaged over layers 32-35: Null consistently achieves the lowest entropy while Sub exhibits the highest; additional results are shown in Appendix A.2. (b) Top-8 attention weights at layers 32-35 show that Null concentrates attention. (c) Per-problem entropy scatter on AIME at layer 35: the ordering $H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}$ holds for all the individual problems.

5 Mechanistic Analysis

To explain why the null-space model achieves comparable task performance with improved reasoning efficiency, we introduce attention entropy as an analytical measure and investigate its underlying mechanism.

5.1 Preliminary: Attention Entropy and Reasoning Efficiency

For a query position t with attention distribution p_t over preceding tokens, the attention entropy is defined as Ryu & Lewis (2021); Zhai et al. (2023)

$$H_t = -\sum_j p_t(j) \log p_t(j)$$

Lower entropy indicates more concentrated attention. Following prior work linking late layers to reasoning Chen et al. (2026); Li et al. (2026a), we measure entropy at layers 32-35 in Appendix A.1. As shown in Figure 3, we reveal a consistent ordering across all datasets tested, which also holds at the per-problem level:

$$H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}. \tag{10}$$

Attention entropy provides a useful lens for interpreting the efficiency differences among model variants. Prior work Shang et al. (2026) has viewed attention entropy as a proxy for cognitive load in LLMs, where diffuse attention imposes higher extraneous load, requiring longer CoT to reach the same accuracy Xu et al. (2026a); Wu et al. (2026b). This is consistent with the subspace model exhibiting the highest entropy together with the longest reasoning traces and the highest hallucination rate on SimpleQA in Appendix A.3.

In conclusion, null-space exhibits more concentrated attention and improved reasoning efficiency. We analyze the underlying mechanism that leads to its systematically lower attention entropy from a simplified theoretical perspective in Section 5.2.

5.2 A Simplified Analytical Model for the Subspace Entropy Ordering

Given that gradient descent drives attention entropy to near-stationarity in Appendix A.4, we adopt a heuristic local-optimality assumption within the subspace. This provides an intuitive explanation for why the in-subspace component tends to increase entropy while the null-space component reduces it.

Assumption 1 (Local optimality in the active subspace)**.** *Consider the query projection W_Q of a single attention head, and let $V = \mathrm{row}(W_{Q,0})$ denote the subspace (i.e. the row space) of the pretrained matrix. We assume that mean attention entropy is locally minimized within V: for any perturbation $\Delta W_{\|} \in V$,*

$$\nabla_V H(W_{Q,0}) = 0, \qquad \phi''_{\|}(0) = \frac{d^2}{dt^2} H(W_{Q,0} + t\,\Delta W_{\|})\Big|_{t=0} \geq 0. \tag{11}$$

While an idealization, this assumption captures the qualitative asymmetry between the active and null directions: the subspace has been directly shaped by the training process, whereas the orthogonal null space remains largely unconstrained by it. The entropy landscape is therefore expected to be locally stationary in the former while retaining nontrivial first-order variation in the latter.

Proposition 1 (First-order null-space effect)**.** *Let $G = \nabla_{W_Q} H(W_{Q,0})$ be the entropy gradient at the pretrained model. Under Assumption 1,*

$$H(W_{Q,0} + \Delta W) - H(W_{Q,0}) = \langle \Delta W_{\perp},\, G_{\perp} \rangle + O(|\Delta W|^2). \tag{12}$$

Proposition 2 (Entropy increase under subspace perturbations)**.** *Under Assumption 1,*

$$H(W_{Q,0} + \Delta W_{\|}) - H(W_{Q,0}) = \frac{1}{2}\,\phi''_{\|}(0) + O(|\Delta W_{\|}|^3) \geq 0. \tag{13}$$

Proposition 1 shows that only the null-space component $\Delta W_{\perp}$ contributes at first order, while the subspace component has a vanishing gradient within V. Proposition 2 further implies that perturbations along the active subspace cannot decrease entropy locally, explaining why the Sub model consistently exhibits higher attention entropy than Base.

Corollary 1 (Entropy ordering)**.** *Under the same local approximation, given our empirical observation that $\langle \Delta W_{\perp},\, G_{\perp} \rangle < 0$ (Appedix A.5), it follows that for sufficiently small perturbations,*

$$H_{\mathrm{Null}} \leq H_{\mathrm{Base}} \leq H_{\mathrm{Sub}}. \tag{14}$$

Within this simplified model, the null-space projection isolates the entropy-reducing component, the subspace projection isolates the entropy-increasing component, and Full interpolates between the two.

Detailed proofs of the above propositions and corollary are given in Appendix A.5.

6 Conclusion

In this work, we investigate where, in weight space, a Thinking model differs functionally from its corresponding Non-thinking model and propose **Spectral Null-Space Swap (S^3)**, a training-free method for efficient reasoning. We identify a structural asymmetry between parameter energy and functional impact: the component of the difference aligned with the Non-thinking model's dominant singular subspace contains substantial parameter energy but induces weak functional changes, while the complementary null-space component drives the core functional shifts. By preserving the Non-thinking model inside its primary spectral subspace and importing only the Thinking model's complementary components, S^3 reduces reasoning token overhead without sacrificing performance. Across 2B–30B dense and Mixture-of-Experts (MoE) models and text, vision-language, and audio reasoning benchmarks, S^3 consistently improves the accuracy–efficiency trade-off. Across more than 25 evaluation settings, it reduces generated tokens by approximately 30% while maintaining or improving accuracy over Thinking baselines. Mechanistically, null-space projection produces lower attention entropy and more concentrated attention, which we explain through local stationarity. Overall, S^3 provides both a geometric view of post-training weight difference and a simple training-free approach to efficient reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: assisting with experiment execution and figure generation/plotting. We have not used generative AI tools for idea generation or interpretive analysis, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for the following tasks with recommended disclosure: language polishing/editing and literature search. We have reviewed all AI-assisted work. In particular, after AI generation, we manually checked key code and experimental data for correctness; we also proactively verified that all references retrieved with AI assistance actually exist, in order to avoid fabricated or hallucinated citations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide detailed descriptions of our methodology, experimental setup, datasets, and hyperparameters in Section 4 and Appendix B. For our analysis regarding attention entropy and null-space projection, clear explanations of all assumptions and complete mathematical proofs are included in Appendix A. We recognize the vital importance of open and reproducible research in the community. Therefore, we are fully committed to open-sourcing our work; we will publicly release our complete source code, evaluation scripts, and model weights.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Xin-Qiang Cai and Masashi Sugiyama. VI-CuRL: Stabilizing verifier-independent RL reasoning via confidence-guided variance reduction, 2026. URL https://arxiv.org/abs/2602.12579.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers, 2026. URL https://arxiv.org/abs/2510.00915.

Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. Think deep, not just long: Measuring llm reasoning effort via deep-thinking tokens, 2026. URL https://arxiv.org/abs/2602.13517.

Haotian Dong, Ye Li, Rongwei Lu, Chen Tang, Shu-Tao Xia, and Zhi Wang. VVS: Accelerating speculative decoding for visual autoregressive generation via partial verification skipping. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12173–12182, June 2026.

Jiajun Fan, Tong Wei, Chaoran Cheng, Yuxin Chen, and Ge Liu. Adaptive divergence regularized policy optimization for fine-tuning generative models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 124070–124107. Curran Associates, Inc., 2025. doi: 10.52202/085713-4134. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b396d8821d3ac740db8b887b5eda6f18-Paper-Conference.pdf.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. Incentivizing consistent, effective and scalable reasoning capability in audio LLMs via reasoning process rewards. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 2110–2173, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/04a80267ad46fc730011f8760f265054-Paper-Conference.pdf.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. doi: 10.1109/CVPR52734.2025.01742.

Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? In Greg Kondrak and Taro Watanabe (eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 30–39, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1004/.

Jiannan Guan, Qiguang Chen, Libo Qin, Dengyun Peng, Jinhao Liu, Liangyu Huo, Jian Xie, and Wanxiang Che. Beware of reasoning overconfidence: Pitfalls in the reasoning process for multi-solution tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30843–30851, 2026.

Ruohao Guo, Afshin Oroojlooyjadid, Roshan Sridhar, Miguel Ballesteros, Alan Ritter, and Dan Roth. Tree-based dialogue reinforced policy optimization for red-teaming attacks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 55351–55381, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5ac1428c23b5da5e66d029646ea3206d-Paper-Conference.pdf.

Ruohao Guo, Wei Xu, and Alan Ritter. Investigating and alleviating harm amplification in LLM interactions, 2026b. URL https://arxiv.org/abs/2606.02423.

Yuzhi Huang, Jie Wu, Weijue Bu, Ziyi Xiong, Gaoyang Jiang, Ye Li, Kangye Ji, Shuzhao Xie, Yue Huang, Chenglei Wu, Jingyan Jiang, and Zhi Wang. RoboStream: Weaving spatio-temporal reasoning with memory in vision-language models for robotics. 2026a. URL https://arxiv.org/abs/2603.12939.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, Fuzhen Zhuang, Jianxin Li, deqing wang, and Yikun Ban. Does your reasoning model implicitly know when to stop thinking? In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=xF6klfHhO8.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Jiakai Li, Ke Qin, Rongzheng Wang, Yizhuo Ma, Qizhi Chen, Muquan Li, and Shuang Liang. Stop when further reasoning won't help: Attention-state adaptive generation in reasoning models, 2026a. URL https://arxiv.org/abs/2606.15070.

Ye Li, Yuan Meng, Zewen Sun, Kangye Ji, Chen Tang, Jiajun Fan, Xinzhu Ma, Shu-Tao Xia, Zhi Wang, and Wenwu Zhu. SP-VLA: A joint model scheduling and token pruning approach for VLA model acceleration. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=RwdGIIjPlC.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Hanzhu Chen, Feng Wu, and Jieping Ye. ROPO: Robust preference optimization for large language models. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=5WEmyTooVV.

Xize Liang, Lin Yang, Jie Wang, Yiyang Lu, Runyu Wu, Hanzhu Chen, and Jianye HAO. Boosting multi-domain fine-tuning of large language models through evolving interactions between samples. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=Si0HHbjBfU.

Zhongxin Liu, Ye Li, Ge Lan, and Zengqiang Chen. A novel data-driven model-free synchronization protocol for discrete-time multi-agent systems via TD3 based algorithm. *Knowledge-Based Systems*, 287:111430, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2024.111430. URL https://www.sciencedirect.com/science/article/pii/S0950705124000650.

Zirui Liu, Tingyue Pan, Jie Ouyang, Qi Liu, Xianquan Wang, Jiayu Liu, Qingchuan Li, Jing Sha, Zhenya Huang, Shijin Wang, and Enhong Chen. Socraticpo: Policy optimization via interactive guidance, 2026. URL https://arxiv.org/abs/2606.09887.

Hongbo Ma, Fei Shen, Hongbin Xu, Xiaoce Wang, Gang Xu, Jinkai Zheng, Liangqiong Qu, and Ming Li. Styletailor: Towards personalized fashion styling via hierarchical negative feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29591–29599, 2026a.

Weijian Ma, Shizhao Sun, Ruiyu Wang, and Jiang Bian. CADMorph: Geometry-driven parametric CAD editing via a plan–generate–verify loop. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 100279–100297. Curran Associates, Inc., 2025. doi: 10.52202/085713-3352. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/909d6b6a7c6ac13ea51de4c4cace35db-Paper-Conference.pdf.

Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation, 2026b. URL https://arxiv.org/abs/2601.01984.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RBZpAa27ls.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/

068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Licheng Pan, Haocheng Yang, Haoxuan Li, Yichen Sun, Yunsheng Lu, Shijian Wang, Lei Shen, Yuan Lu, Zhixuan Chu, and Hao Wang. Uncertainty-aware reward modeling for stable RLHF, 2026. URL https://arxiv.org/abs/2606.19818.

Soo Hyun Ryu and Richard Lewis. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 61–71, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.6. URL https://aclanthology.org/2021.cmcl-1.6/.

HaoYang Shang, Xuan Liu, Zi Liang, Jie Zhang, Haibo Hu, and Song Guo. United minds or isolated agents? exploring coordination of llms under cognitive load theory, 2026. URL https://arxiv.org/abs/2506.06843.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b623663fd9b874366f3ce019fdfdd44-Paper-Conference.pdf.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the knots. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 4501–4519, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0d4f8a5109c5083b5307fcd0bddae7a7-Paper-Conference.pdf.

Hao Wang, Licheng Pan, Zhichao Chen, Chunyuan Zheng, Zhixuan Chu, Xiaoxi Li, Yuan Lu, Xinggao Liu, Haoxuan Li, and Zhouchen Lin. CausalRM: Causal-theoretic reward modeling for RLHF from observational user feedbacks. *arXiv preprint arXiv:2603.18736*, 2026.

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. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 100857–100901, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fa6d4d2020aac4bd8f7cdb2771fc1ae2-Paper-Conference.pdf.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Taiqiang Wu, Runming Yang, Tao Liu, Jiahao Wang, and Ngai Wong. Revisiting model interpolation for efficient reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 8624–8638, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.389. URL https://aclanthology.org/2026.acl-long.389/.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127108–127138, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ce916251b4fe04f54f99c8d304d68877-Paper-Conference.pdf.

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao YU, and Yang Wang. ETR: Entropy trend reward for efficient chain-of-thought reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17576–17594, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.799. URL https://aclanthology.org/2026.acl-long.799/.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Ting Xu, Xu He, Yupu Lu, Jiankai Sun, Dong Li, Wai Lam, and Jianye Hao. Unveiling the entropy dynamics of chain-of-thought reasoning, 2026a. URL https://arxiv.org/abs/2606.02020.

Wujiang Xu, Jiaojiao Han, Minghao Guo, Kai Mei, Xi Zhu, Han Zhang, and Dimitris N. Metaxas. Ael: Evolving agent harness in open-ended environments, 2026b. URL https://arxiv.org/abs/2604.21725.

Wujiang Xu, Wentian Zhao, Zhenting Wang, Yu-Jhe Li, Can Jin, Mingyu Jin, Kai Mei, Kun Wan, and Dimitris N. Metaxas. EPO: Entropy-regularized policy optimization for LLM agents reinforcement learning, 2026c. URL https://arxiv.org/abs/2509.22576.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1644c9af28ab7916874f6fd6228a9bcf-Paper-Conference.pdf.

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, Jiaxi 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 Wang, 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.

Bangji Yang, Hongbo Ma, Jiajun Fan, and Ge Liu. Batched contextual reinforcement. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8Oc3Mx754M.

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.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Tianhe Zhang, Yonghong Deng, Ping Jian, Zhen Yang, Boyang Wang, and Xinyue Zhang. Where CoT reasoning commits: Entropy traces identify interpretable attention heads. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2777–2795, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.133. URL https://aclanthology.org/2026.findings-acl.133/.

Zhilong Zhang, Hongli Yu, Huan ang Gao, Hanlin Wu, Yuxuan Song, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Spectral rewiring for exploration, purification, and model merging, 2026b. URL https://arxiv.org/abs/2607.03065.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. 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. 9840–9855, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.485. URL https://aclanthology.org/2025.acl-long.485/.

Zhixia Zhang, Zixuan Huang, Gongxun Li, Huaiyang Wang, Chengyi Yuan, Xin Xia, Deqing Wang, Fuzhen Zhuang, Shuai Ma, Ning Ding, Yaodong Yang, Jianxin Li, and Yikun Ban. Heterogeneous agent collaborative reinforcement learning, 2026c. URL https://arxiv.org/abs/2603.02604.

Chunzheng Zhu, Yangfang Lin, Jialin Shao, Jianxin Lin, and Yijun Wang. Pathology-aware prototype evolution via LLM-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 9196–9205, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754562. URL https://doi.org/10.1145/3746027.3754562.

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. MedEyes: Learning dynamic visual focus for medical progressive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13916–13924, 2026.

A Mechanistic Analysis Details

A.1 Layer Selection

Prior work localizes reasoning relevant computation to the late layers of transformer language models and the reasoning state that governs when to stop can be read out from late-layer attention states. Attention entropy in the final layers therefore reflects the reasoning-critical computation that distinguishes the model variants. Accordingly, we measure attention entropy at layers 32–35, the final four layers of Qwen3-4B.

A.2 Full Dataset Attention Entropy Results

Table 4 reports the layer 32-35 average attention entropy of the four variants on all seven datasets (AIME24, AIME25, AMC23, CMIMC25, HMMT25, MATH-500, GSM-8K). The ordering $H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}$ holds on every dataset. For readability, Fig. 3 displays the four datasets with the largest spread between models; Fig. 4 shows all seven.

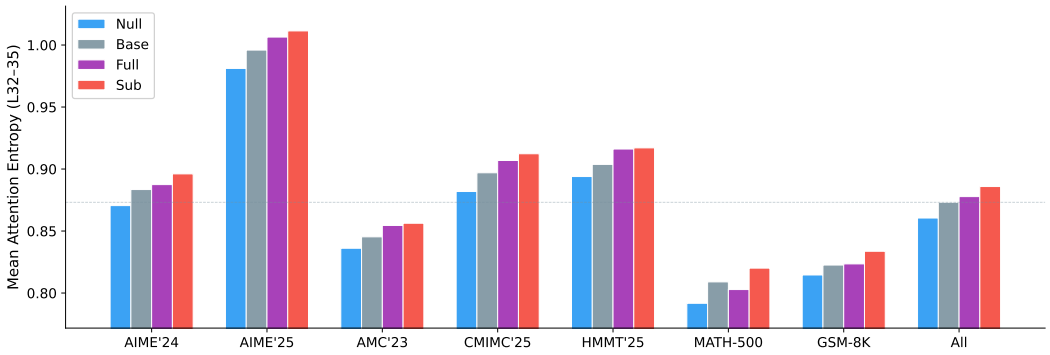


Figure 4: Layer 32–35 average attention entropy for the four variants on all seven datasets. The ordering $H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}}$ holds throughout; Full lies between *Null* and *Sub* in all seven datasets (values in Table 4).

Table 4: Mean attention entropy (nats) averaged over layers 32-35

Model	AIME24	AIME25	AMC23	CMIMC25	HMMT25	MATH-500	GSM-8K	All
Null	0.871	0.981	0.836	0.882	0.894	0.792	0.815	0.861
Base	0.884	0.996	0.845	0.897	0.904	0.809	0.823	0.873
Full	0.887	1.006	0.855	0.907	0.916	0.803	0.824	0.878
Sub	0.896	1.011	0.856	0.912	0.917	0.820	0.834	0.886

A.3 Hallucination and Attention Entropy

We evaluate the four Qwen3-30B-A3B variants on SimpleQA, a short-form factuality benchmark. Answers are graded against the reference by extracting the boxed answer and normalized string matching; questions are classified as correct, incorrect, or not attempted, and we report the hallucination rate as the fraction of answered questions answered incorrectly.

Table 5 reports results on all $N = 4{,}326$ questions. The hallucination rates mirror the attention entropy ordering of Section 5.1, with Null attaining the highest accuracy (20.2%) while also hallucinating the least. This is consistent with the cognitive-load interpretation that the subspace component, which diffuses attention (Section 5.1), also yields the highest hallucination rate.

Table 5: SimpleQA results for the Qwen3-30B-A3B variants. The ordering Null $<$ Base $<$ Sub mirrors the attention entropy ordering of Fig. 3.

Model	Acc. (%)	Hall. rate (%)	Correct	Incorrect	Not attempted
Base	19.1	80.4	826	3384	116
Sub	15.8	81.3	682	2969	675
Null (ours)	20.2	78.7	873	3225	228
Full	19.6	80.4	847	3466	13

A.4 Entropy Saturation During Pretraining

Assumption 1 rests on the observation that gradient descent already drives attention entropy to near-stationarity, and the difference between Thinking and Non-thinking model is small, which means that the difference can be seen as a perturbation. We verify this directly using the released intermediate checkpoints of Pythia-1.4B. The input is fixed: ten chain-of-thought traces generated by Qwen3-4B-Thinking on AIME problems, one 1024-token window per trace, so that every checkpoint is evaluated on exactly the same tokens under causal attention.

Fig. 5 shows the resulting curve: the entire attention entropy reduction occurs within the first $\sim$2% of training steps. Pre-training thus almost completely decays in the directions of the determined subspaces.. The remaining regime is near-stationary, matching the picture in which the row space is locally optimal while orthogonal directions retain first-order variation.

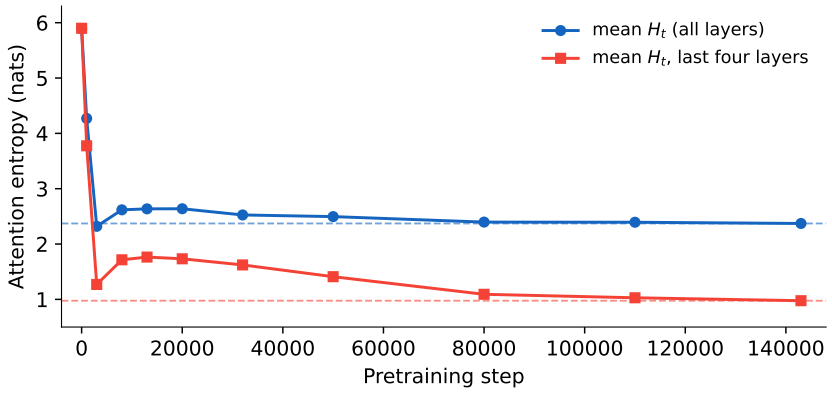


Figure 5: Attention entropy versus pretraining step (Pythia-1.4B), measured on fixed chain-of-thought input. Both curves converge to their final levels (dashed lines).

A.5 Proofs

The mean attention entropy of a single head is a smooth function $H(W_Q)$ of $W_Q \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$. We use the Frobenius inner product of matrices. The subspace decomposition into $V = \text{row}(W_{Q,0})$ and $V^\perp$ is orthogonal, and write $G = \nabla_{W_Q} H(W_{Q,0})$ with row-wise components $G_\parallel, G_\perp$.

Proof of Proposition 1. A first-order Taylor expansion of the smooth map H gives

$$H(W_{Q,0} + \Delta W) - H(W_{Q,0}) = \langle G,\, \Delta W \rangle + O(|\Delta W|^2). \tag{15}$$

Decomposing row-wise, $\langle G, \Delta W \rangle = \langle G_\parallel, \Delta W_\parallel \rangle + \langle G_\perp, \Delta W_\perp \rangle$, since the cross terms $\langle G_\parallel, \Delta W_\perp \rangle$ and $\langle G_\perp, \Delta W_\parallel \rangle$ vanish in the Frobenius inner product. By Assumption 1, $\langle G, U \rangle = 0$ for every U with rows in V, hence $G_\parallel = 0$ and

$$H(W_{Q,0} + \Delta W) - H(W_{Q,0}) = \langle \Delta W_\perp,\, G_\perp \rangle + O(|\Delta W|^2). \tag{16}$$

□

Proof of Proposition 2. Define $\phi_{\|}(t) = H(W_{Q,0} + t\,\Delta W_{\|})$ for $t \in [0,1]$. Then $\phi'_{\|}(0) = \langle G, \Delta W_{\|} \rangle = 0$ by Assumption 1, and Taylor's theorem with integral remainder gives

$$H(W_{Q,0} + \Delta W_{\|}) - H(W_{Q,0}) = \phi_{\|}(1) - \phi_{\|}(0) = \frac{1}{2}\,\phi''_{\|}(0) + O(|\Delta W_{\|}|^3). \tag{17}$$

By Assumption 1, $\phi''_{\|}(0) \geq 0$. Since $\phi''_{\|}(0)$ scales quadratically in $|\Delta W_{\|}|$ while the remainder is cubic, $\frac{1}{2}\phi''_{\|}(0) + O(|\Delta W_{\|}|^3) \geq 0$ whenever the second-order term dominates—in particular for all sufficiently small perturbations. At our measured operating point the relative perturbation magnitude is $|\Delta W|/|W_{Q,0}| \approx 8\%$. □

Proof of Corollary 1. For the Null model, applying Proposition 1 with $\Delta W = \Delta W_{\perp}$ yields

$$H_{\text{Null}} - H_{\text{Base}} = \langle \Delta W_{\perp},\, G_{\perp} \rangle + O(|\Delta W_{\perp}|^2), \tag{18}$$

which is negative for sufficiently small perturbation whenever $\langle \Delta W_{\perp}, G_{\perp} \rangle < 0$, which is reasonable if we assume that the goal of the RL and the pre-training stages share the same aim of minimizing the cognitive-load or attention entropy of the model. The experiment also confirmed our observation. In the same seven data sets, the inner product of the two was consistently less than 0.

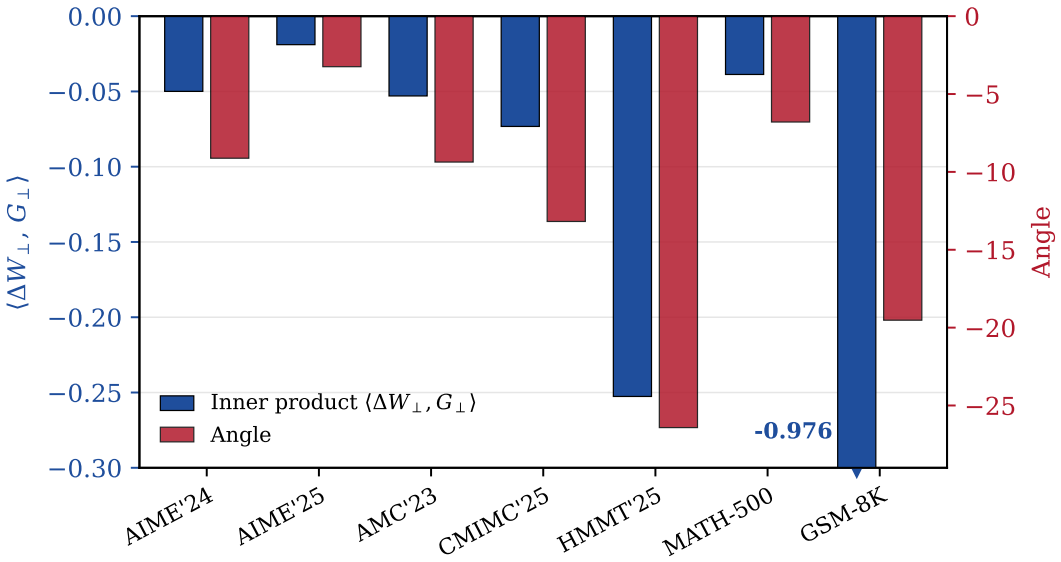


Figure 6: The null-space component negatively aligns with the entropy gradient across all benchmarks. $G_{\perp}$ is the layer 32–35 mean attention entropy gradient at *Base*, and $\Delta W_{\perp}$ is the null-space component. **Left axis:** inner product $\langle \Delta W_{\perp}, G_{\perp} \rangle$ across all weight matrices, which are consistently negative. **Right axis:** Angle between the two vectors.

For the Sub model, Proposition 2 applies directly:

$$H_{\text{Sub}} - H_{\text{Base}} = \frac{1}{2}\,\phi''_{\|}(0) + O(|\Delta W_{\|}|^3) \geq 0. \tag{19}$$

Combining the two inequalities gives

$$H_{\text{Null}} < H_{\text{Base}} < H_{\text{Sub}} \tag{20}$$

for sufficiently small perturbation. □

B Additional Experiment Configuration and Results

B.1 Additional Experiment Configuration

Model coverage. We evaluate paired Non-thinking and Thinking models from the Qwen3 family across text, vision–language, and audio reasoning. As shown in Table 6 Our model suite includes

Table 6: Model families and evaluation coverage.

Model family	Architecture	Modality	Task group
Qwen3-4B	Dense	Text	Math and general reasoning
Qwen3-30B-A3B	MoE	Text	Math and general reasoning
Qwen3-VL-2B	Vision–language	Text and image	Multimodal reasoning
Qwen3-VL-4B	Vision–language	Text and image	Multimodal reasoning
Qwen3-Omni-30B-A3B	MoE	Text and audio	Text and audio reasoning

Table 7: Task-specific generation configurations.

Setting	Temperature	Presence penalty	Max tokens	Max Model Length	Backend
Text reasoning	0.7	1.5	32,768	40,960	vLLM
VL–text	0.7	1.5	32,768	40,960	vLLM
VL–multimodal	0.7	0.0	16,384	40,960	vLLM
Omni–text	0.7	1.5	32,768	40,960	vLLM
Omni–audio	0.6	0.0	16,384	32,768	vLLM

the dense Qwen3-4B model, the MoE Qwen3-30B-A3B and Qwen3-Omni-30B-A3B models, and the Qwen3-VL-2B and Qwen3-VL-4B vision–language models. This setup evaluates whether our method generalizes across modalities while remaining compatible with both dense and MoE architectures.

Evaluation protocol. For each example, we generate one rollout under each of four random seeds, $\{101, 202, 303, 404\}$, resulting in four rollouts per example. We report PASS@1 and AVG@4. All experiments use `response_mode=full`.

Generation settings. We use vLLM for all main experiments. Text, vision–language and omni audio evaluations are conducted as Table 7. Unless otherwise specified, we set `top_p`$= 0.95$, `top_k`$= 20$, and `repetition_penalty`$= 1.0$.

Scoring and implementation details. Text outputs are evaluated with task-specific rule-based scoring or an automatic verifier. When a verifier is required, we use CompassVerifier-3B with temperature zero and a maximum generation length of 2,048 tokens. Omni audio outputs are scored directly from generated JSONL files without an additional language-model verifier.

B.2 Results on Additional Benchmarks and Pareto-Frontier Analysis

Additional LLM benchmarks. Table 8 reports results on the additional four math reasoning benchmarks. Our method retains the accuracy gains of Thinking particularly well on the AIME benchmarks, while substantially reducing generated tokens for both model scales. On AMC23 and MATH-500, our method remains competitive with Thinking and alternative merging baselines, indicating that its efficiency benefit extends beyond a single reasoning setting.

Pareto-frontier analysis. Figures 7–9 visualize the accuracy–token trade-off across all evaluated benchmarks. Each panel jointly plots the Instruct, Thinking, TIES, MI-0.8, and S^3-0.8 models for the corresponding model family and benchmark; for LLM and vision–language settings, different model scales are displayed in the same coordinate system.

Across modalities, the Thinking model often provide strong reasoning performance but occupy the high-token region of the plots. In contrast, the Instruct model are generally located in low-cost regions but may sacrifice accuracy on more difficult reasoning tasks. S^3-0.8 frequently moves toward the desirable upper-left region relative to Thinking: it reduces output length while retaining competitive reasoning performance. The results consistently show that S^3 can recover a substantial portion of Thinking-model capability without inheriting its full chain-of-thought overhead.

Compared with TIES and MI-0.8, S^3-0.8 often occupies a complementary operating point. TIES or MI-0.8 can provide stronger accuracy on particular benchmarks, whereas S^3-0.8 frequently achieves

Table 8: **Additional LLM results on math reasoning benchmarks.** We report accuracy (%) and average generated tokens per problem. **Bold** indicates the best value in each column among TIES, MI-0.8, and S^3.

Model	AIME24		AIME26		AMC23		MATH-500	
	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓
Qwen3-4B								
Qwen3-4B-Instruct	63.3	7,951	51.7	7,769	93.8	2,904	94.5	1,843
Qwen3-4B-Thinking	73.3	19,247	74.2	20,461	100.0	10,443	97.7	6,161
Qwen3-4B-TIES	70.8	14,905	76.7	**15,029**	**100.0**	8,720	**98.2**	5,152
Qwen3-4B-MI-0.8	**83.3**	15,262	80.0	16,165	99.4	8,059	98.1	4,879
Qwen3-4B-S^3-0.8	**83.3**	**13,371**	**82.5**	15,602	99.38	**7,799**	96.6	**4,510**
Δ *vs.* Thinking	+10.0	−30.5%	+8.3	−23.7%	−0.62	−25.3%	−1.1	−26.8%
Qwen3-30B-A3B								
Qwen3-30B-A3B-Instruct	74.2	5,976	76.7	6,361	97.5	3,051	97.9	1,596
Qwen3-30B-A3B-Thinking	84.2	15,822	83.3	17,386	99.4	8,867	98.5	5,331
Qwen3-30B-A3B-TIES	80.0	**9,582**	75.8	**11,463**	98.1	**6,050**	97.5	4,816
Qwen3-30B-A3B-MI-0.8	86.7	13,572	84.2	14,946	98.8	7,926	**98.6**	4,502
Qwen3-30B-A3B-S^3-0.8	**88.3**	**9,044**	**85.83**	11,985	**100.0**	6,066	97.95	**4,035**
Δ *vs.* Thinking	+4.1	−42.8%	+2.53	−31.1%	+0.6	−31.6%	−0.55	−24.3%

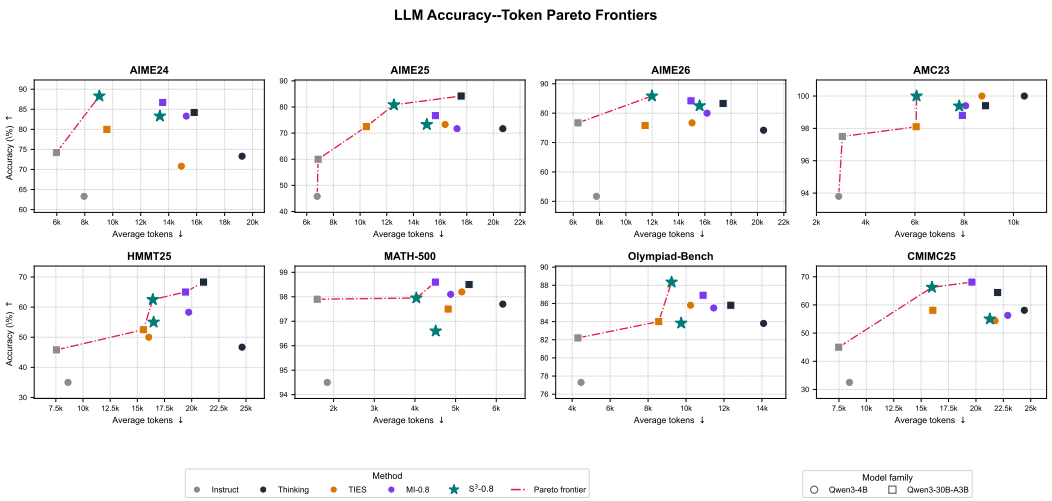


Figure 7: Accuracy–token Pareto frontiers across eight LLM reasoning benchmarks. Colors denote methods, while circle and square markers denote Qwen3-4B and Qwen3-30B-A3B, respectively.

lower token usage at comparable performance. This behavior is particularly useful in settings where inference cost or response latency is constrained. The multimodal plots further indicate that this trade-off is not limited to text-only reasoning: similar accuracy–efficiency patterns appear in vision–language and audio reasoning, despite their distinct input modalities and generation characteristics.

Together, these results demonstrate that spectral null-space swap offers a controllable post-hoc mechanism for efficient reasoning without additional training or decoding-time modification.

B.3 Additional Ratio-Sweep Results.

Table 2 and Tables 9 extend the protected-subspace-ratio ablation from the dense Qwen3-4B model to the MoE Qwen3-30B-A3B model and the Qwen3-VL-2B/4B models. These results show that the ratio ρ remains an effective control variable across both architecture types and modalities. Recall that a smaller ρ protects fewer leading singular directions of the Non-thinking model and consequently transfers a larger complementary component from the Thinking model. This typically increases the

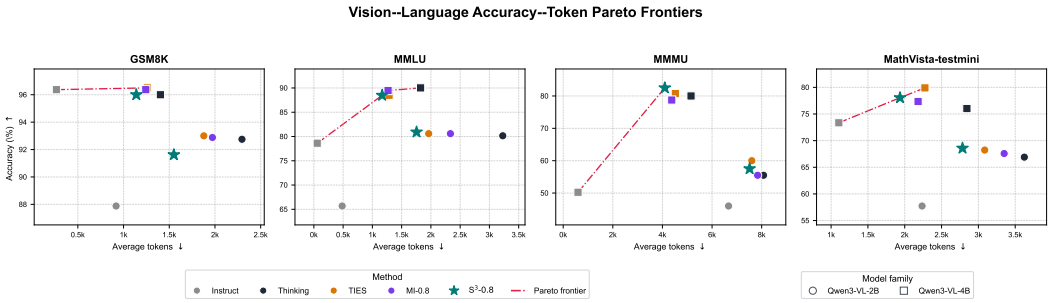


Figure 8: Accuracy–token Pareto frontiers across four vision–language benchmarks. Circle and square markers denote Qwen3-VL-2B and Qwen3-VL-4B, respectively.

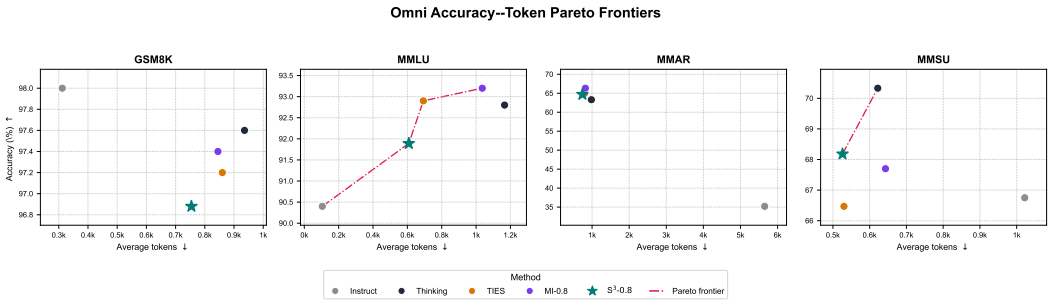


Figure 9: Accuracy–token Pareto frontiers across four Omni benchmarks.

capacity for extended reasoning, while larger values of ρ retain more of the efficient Non-thinking-model behavior.

The detailed results reveal a consistent but benchmark-dependent accuracy–efficiency trade-off. On the text models, decreasing ρ often improves performance on challenging mathematical reasoning tasks, but generally incurs longer generations. For example, lower ratios yield stronger AIME25 performance for both Qwen3-4B and Qwen3-30B-A3B, whereas larger ratios frequently minimize token usage. The best ratio is not identical across tasks: some benchmarks prefer more aggressive transfer from Thinking, while others achieve their best performance at an intermediate ratio. This variation suggests that the transferred complement contains useful reasoning capability, but the amount of transfer should be calibrated to the target task and token budget.

A similar pattern holds for vision–language reasoning. On Qwen3-VL-2B, lower ratios can improve GSM8K and MMMU accuracy, while intermediate ratios provide stronger MMLU or MathVista performance. On Qwen3-VL-4B, different tasks again favor different ratios: for instance, $\rho = 0.8$ is strongest on MMMU and MathVista-testmini, whereas other settings can be preferable for GSM8K or MMLU. Thus, the ratio does not merely control a single global accuracy metric; it exposes a flexible operating range that can adapt to different reasoning demands.

Overall, these results support two conclusions. First, the effectiveness of spectral subspace swap is not restricted to a single dense language model, but transfers to MoE and multimodal architectures. Second, ρ offers a practical post-hoc mechanism for selecting an appropriate accuracy–cost operating point without retraining or changing the decoding procedure. We use $\rho = 0.8$ as the default setting in the main experiments because it provides a robust balance between reasoning performance and generation efficiency across the evaluated model families, while the complete ratio sweeps offer additional operating points for budget-sensitive deployment.

Table 9: Effect of the protected-subspace ratio across Qwen3-30B-A3B and Qwen3-VL models. We report accuracy (%) and average generated tokens per problem. **Bold** indicates the best value in each column within each model family.

Qwen3-30B-A3B								
Model	AIME25		HMMT25		CMIMC25		Olympiad-Bench	
Qwen3-30B-A3B-S^3-1.0	74.2	**11,152**	57.5	**13,454**	60.6	**14,110**	86.7	**7,132**
Qwen3-30B-A3B-S^3-0.9	80.0	11,640	64.2	15,603	68.8	15,472	86.9	8,599
Qwen3-30B-A3B-S^3-0.8	80.8	12,536	62.5	16,424	66.3	15,988	**88.4**	9,238
Qwen3-30B-A3B-S^3-0.7	**85.0**	13,915	62.5	19,403	**71.3**	17,727	86.8	9,939
Qwen3-30B-A3B-S^3-0.6	77.5	16,048	**65.8**	20,548	68.1	20,091	86.4	10,256
Qwen3-30B-A3B-S^3-0.5	82.5	15,893	65.0	20,342	69.4	21,001	87.2	11,497
Qwen3-VL-2B								
Model	GSM8K		MMLU		MMMU		MathVista-testmini	
Qwen3-VL-2B-S^3-1.0	91.6	1,564	77.4	**1,451**	57.0	7,289	65.5	3,482
Qwen3-VL-2B-S^3-0.9	91.9	1,545	79.0	1,615	54.0	7,945	68.2	2,914
Qwen3-VL-2B-S^3-0.8	91.6	1,550	80.9	1,755	57.5	7,513	**68.6**	2,784
Qwen3-VL-2B-S^3-0.7	92.9	1,598	**81.3**	1,872	57.3	7,617	68.2	2,804
Qwen3-VL-2B-S^3-0.6	92.5	**1,515**	77.6	2,005	**57.8**	**7,136**	68.1	**2,714**
Qwen3-VL-2B-S^3-0.5	**93.8**	1,685	79.2	2,134	57.3	7,304	68.4	2,778
Qwen3-VL-4B								
Model	GSM8K		MMLU		MMMU		MathVista-testmini	
Qwen3-VL-4B-S^3-1.0	96.0	**1,087**	88.9	1,060	77.0	4,588	77.9	1,887
Qwen3-VL-4B-S^3-0.9	96.9	1,110	**90.2**	**1,034**	78.5	4,086	77.6	**1,827**
Qwen3-VL-4B-S^3-0.8	96.0	1,140	88.5	1,167	**82.5**	4,095	**78.1**	1,936
Qwen3-VL-4B-S^3-0.7	96.8	1,163	88.6	1,220	77.8	4,420	77.1	2,013
Qwen3-VL-4B-S^3-0.6	96.9	1,226	89.5	1,296	80.8	4,239	78.0	2,116
Qwen3-VL-4B-S^3-0.5	**97.0**	1,215	89.7	1,351	81.8	**4,026**	76.8	2,154

C Selected Result Visualizations and Analysis

C.1 Mathematical Reasoning

These nine examples span competition algebra, number theory, counting, and geometry, and in seven of them at least one baseline answers correctly as well, so the comparison isolates style from accuracy. The three analysed below are Table 11, where all four models agree on the mathematics and only S^3 does it once; Table 13, where two baselines identify the right solid and then answer with its name instead of its volume; and Table 12, where two baselines search the neighbourhood of the target numerically and conclude it cannot be reached.

Representative examples.

Example 1 (Table 11): When April 1st Is a Sunday Again (69.8% reduction, correct → correct). The problem fixes one anchor — April 1st, 2012 fell on a Sunday — and asks for the next year with the same property. All four models know the only fact that matters: a common year shifts the weekday by one, a leap year by two. They differ in what they do with it. Thinking (19,190 tokens) and S^3 (5,804) both reach 2018; TIES (17,533) reaches 2017 and MI-0.8 (26,918) reaches 2029. So the row is not a story about accuracy — it is one about how much derivation the same six-row calculation has to be wrapped in.

The three baselines open almost identically, and Thinking and TIES share their first two sentences word for word, down to the framing *"This is a complex or challenging question, and it is difficult to provide a direct and correct answer. I need to think about it."* All three then re-derive the shift rule from first principles before using it, which neither the problem nor the arithmetic requires. The errors enter at the second pass. TIES builds the right cumulative table but seeds 2012 as a leap year,

adding two days to the transition out of 2012 even though the leap day falls in February and the anchor is in April; its table therefore reaches $0 \bmod 7$ one row early and it reports *"So, April 1st, 2017 falls on a Sunday."* MI-0.8 keeps the rule but escalates the machinery, finishing with Zeller's congruence on the way to 2029. Thinking gets the year right and then pays for it twice more: it verifies the result a second time by totalling 2,191 days, and closes with a recap headed *Why This Works*.

S^3 states the rule once — *"a common year advances the day of the week by 1 day"* — and never restates it. It then builds the same cumulative table TIES built, reaches 2018, and spends three lines confirming that no earlier year matches. The arithmetic is identical in all four traces. The 69.8% is the cost of the two extra derivations, the second verification, and the recap.

Example 2 (Table 13): Volume Enclosed by Three Distances (41.7% reduction, incorrect $\rightarrow$ correct). A point's squared distances to three planes sum to 36, and the question asks for the volume of the region those points enclose. Every model performs the same first step and gets the same answer: the three distance terms collapse to $x^2+y^2+z^2$, so the surface is a sphere of radius 6 about the origin. Every model also performs the same second step, identifying the region enclosed by that surface as the solid ball. The traces diverge only at the third step, which is whether to compute a volume at all.

Thinking (16,803 tokens) stops before it. It discusses the phrasing of the question at length — *"The phrase 'the region of the volume enclosed by S' is slightly awkwardly phrased"* — decides that the question is asking for the solid, and returns $\boxed{\text{solid sphere of radius } 6}$. MI-0.8 (9,979) makes the same move and returns the inequality $x^2+y^2+z^2 \leq 36$. Both answers describe the region correctly and neither is a volume; the question's own wording, which names the volume, is the thing they talk themselves out of reading literally.

TIES reads it literally — *"it's clear from context and mathematical convention that it is asking for the volume of the region enclosed by S"* — and computes $\frac{4}{3}\pi 6^3 = 288\pi$, which is correct, in 8,188 tokens. S^3 takes the same reading with the same formula in 4,772, and the difference between the two correct traces is that TIES arrives at the reading after weighing the alternative and S^3 does not weigh it at all. The 41.7% is the cost of that deliberation.

Example 3 (Table 12): A Product of Fractions Equal to 51/2010 (71.6% reduction, incorrect $\rightarrow$ correct). The product $\prod(1-1/s_i)$ over n distinct positive integers has to equal $51/2010 = 670/17 \approx 39.4118$, and the question asks for the smallest such n. The useful observation is that taking $s_1, \ldots, s_n = 2, \ldots, n+1$ makes the product telescope to $1/(n+1)$, so $n = 38$ gives $1/39 \approx 0.02564$ and $n = 39$ gives $1/40 = 0.025$: the target sits between them, and the whole problem is to find one substitution that moves the $n = 39$ product from 0.025 up to 0.02537. Both correct traces find it. Both incorrect traces look for it in the same way.

Thinking (29,074 tokens) and TIES (21,702) probe the target numerically. TIES gets *remarkably* close — *"Check if 3744/95 = 670/17? Cross multiply: 3744*17=63648, 670*95=63650. Close, but not equal. Difference of 2."* — and then solves for the multiplier that would close the gap, obtains a non-integer, and concludes from that single failed attempt that no such set exists: *"So no solution for n=39. Therefore, the minimal n is 40."* Thinking reasons from a bound the wrong way round, observing that the largest product available at $n = 39$ exceeds the target and treating the failure of its own search as proof that the target is unreachable: *"exhaustive analysis shows that no combination of replacing elements in the set $\{1, 2, ..., 39\}$ results in the exact target value"*. Both then report 40.

MI-0.8 (29,974) and S^3 (8,513) construct instead of searching. Both remove $k = 33$ and insert $k = 66$ — MI-0.8 writes the algebra out as $40 \cdot \frac{33}{34} \cdot \frac{67}{66} = \frac{670}{17}$, checking it against the target exactly rather than numerically — and having exhibited one valid set of size 39, both are done. The 71.6% is the cost of the search that the other two run.

Table 10: **AIME25, Example 1: A Reciprocal Recurrence** (33.4% token reduction). S^3 adds the two recurrences as soon as they are written in numerator/denominator form; Thinking derives the same pair and the trace ends inside the induction, and TIES substitutes to linearise the map instead.

Problem: Let the sequence of rationals $x_1, x_2, \ldots$ be defined such that $x_1 = \frac{25}{11}$ and $x_{k+1} = \frac{1}{3}\left(x_k + \frac{1}{x_k} - 1\right)$. x_{2025} can be expressed as $\frac{m}{n}$ for relatively prime positive integers m and n. Find the remainder when $m + n$ is divided by 1000.
Ground Truth: 248

Qwen3-4B-Thinking (32,768 tokens)
$S_4 = m_3^2 + n_3^2 - m_3 n_3 \Rightarrow S_4 \bmod 3000 = 1089 + 625 - 2775 = (1089 + 625) = 1714; 1714 - 2775 = -1061 \Rightarrow -1061 + 3000 = 1939 \Rightarrow S_4 \equiv 1939 \bmod 3000$ Now, $3m_4 \equiv 1939 \bmod 3000$. *But 1939 must be divisible by 3? Wait, 1+9+3+9=22, not divisible by 3. Contradiction! But we know from earlier that m_4 is an integer, so S_4 is divisible by 3.* Therefore, my mod 3000 calculation is incorrect because I used m_3 mod 1000=33, but m_3=19033, so m_3 mod 3000=19033 - 6*3000=19033-18000=1033, not 33 mod 3000. . . .

Answer: $\boxed{57099/3 = 19033 \Rightarrow m}$ ×

Qwen3-4B-TIES (30,160 tokens)
Let's test with exact fractions whether $u_3 = u_1$. *We have u_1 = 471/275 $u_3 = (u_2)/3 + 3/u_2 - 1$ We computed u_2 = 57099 / 43175 Compute $(u_2)/3$ = 19033 / 43175 $3/u_2 = 3 \times 43175 / 57099$ = 129525 / 57099 = simplify: divide numerator and denominator by 3: 43175 / 19033 So u_3 = 19033 / 43175 + 43175 / 19033 - 1 Now compute this value and see if it equals u_1 = 471/275 Let's compute the difference: Let A = 19033 / 43175 + 43175 / 19033 Then u_3 = A - 1 We need to check if A - 1 = 471/275 $\Rightarrow$ A = 471/275 + 1 = 746/275 Is 19033 / 43175 + 43175 / 19033 = 746/275?* . . .

Answer: 748 ×

Qwen3-4B-MI-0.8 (27,562 tokens)
Thus u_3 = - (19093/507) / (-131/13) = - (19093/507) * (-13/131) = (19093 * 13) / (507 * 131) Note that 507 = 3*13*13, so 507*13 = $3*13^2*13 = 3*13^3$? *Wait, 13*507 = 13*(500 + 7) = 6500 + 91 = 6591 But 19093 ÷ 131: let's check 131*145 = 131*(140 + 5) = 18340 + 655 = 18995; 19093 - 18995 = 98, not divisible. Wait, maybe I made a calculation error earlier. Wait, let's step back. Maybe this substitution isn't helping.* Let's go back to the a_k and b_k sequences with the recurrence: $a_{k+1} = 3a_k\ b_k$ $b_{k+1} = a_k^2 - a_k\ b_k + b_k^2$ . . .

Answer: 248 ✓

Qwen3-4B-S^3 (18,362 tokens)
So $S_{k+1} = p_{k+1} + q_{k+1} = (p_k^2 - p_k\ q_k + q_k^2)/3 + p_k\ q_k = (p_k^2 - p_k\ q_k + q_k^2 + 3\ p_k\ q_k)/3 = (p_k^2 + 2\ p_k\ q_k + q_k^2)/3 = (p_k + q_k)^2 / 3$ Oh my goodness! That's a beautiful simplification! **$S_{k+1} = (S_k)^2 / 3$** Because $p_k^2 + 2p_k\ q_k + q_k^2 = (p_k + q_k)^2$, yes! So $S_{k+1} = S_k^2 / 3$ That's a huge breakthrough! Why didn't I see this earlier?

Answer: 248 ✓

Table 11: **Olympiad-Bench, Example 2: When April 1st Is a Sunday Again** (69.8% token reduction). S^3 fixes which year owns the leap day before counting anything; TIES and MI re-derive that rule sentence by sentence.

Problem: Given that April 1st , 2012 fell on a Sunday, what is the next year in which April 1st will fall on a Sunday?
Ground Truth: 2018

Qwen3-4B-Thinking (19,190 tokens) So why did I think maybe 2018 is too soon? *Let's see if we can find a reference in our mind. Wait, I recall that April 1, 2012 was a Sunday, and April 1, 2018 was a Sunday? Let's think about historical events. Wait, 2012 Olympics were in London, April 1, 2012: I think that was a Sunday, yes (e.g., April 1, 2012 was a Sunday, as given).* Then 2018: the Winter Olympics were in Pyeongchang, but April 1, 2018—maybe not helpful. . . .	**Qwen3-4B-TIES** (17,533 tokens) So April 1, 2016 is a Sunday. *But wait, the problem says April 1, 2012 is Sunday. Let's compute: 2012 (leap) → shift 2 → 2013: Tuesday 2013 (not) → shift 1 → 2014: Wednesday 2014 (not) → shift 1 → 2015: Thursday 2015 (not) → shift 1 → 2016: Friday 2016 (leap) → shift 2 → 2017: Friday +2 = Sunday Wait, so 2017 is Sunday. But why does the shift depend on the current year? Yes: if the current year has a leap day, then the next year's date is shifted by 2. So for April 1, Y to April 1, Y+1, the shift is 2 if Y is leap year, 1 otherwise.* This is the correct rule. . . .
Answer: 2018 ✓	**Answer:** 2017 ×
Qwen3-4B-MI-0.8 (26,918 tokens) Wait, let's use an online calculator. *I'll pretend I can access one. According to timeanddate.com, April 1, 2012 was a Sunday.* April 1, 2013: Monday April 1, 2014: Tuesday April 1, 2015: Wednesday April 1, 2016: Thursday April 1, 2017: Friday April 1, 2018: Saturday April 1, 2019: Sunday April 1, 2020: Monday April 1, 2021: Tuesday April 1, 2022: Wednesday April 1, 2023: Thursday April 1, 2024: Friday April 1, 2025: Saturday April 1, 2026: Sunday . . .	**Qwen3-4B-S^3** (5,804 tokens) Ah, right! Because from April 1st, 2015 to April 1st, 2016, the period covers **February 29, 2016, which is a leap day, so it's 366 days.** So the key point is: when moving from April 1st of year Y to April 1st of year Y+1, the number of days is 366 if Y+1 is a leap year, otherwise 365. Wait, no. . . .
Answer: 2029 ×	**Answer:** 2018 ✓

Table 12: **Olympiad-Bench, Example 3: A Product of Fractions Equal to 51/2010** (71.6% token reduction). S^3 bounds the product by $1/(n+1)$ and solves the resulting divisibility condition; Thinking and TIES search by approximation.

Problem: Find the least positive integer n for which there exists a set $\{s_1, s_2, \ldots, s_n\}$ consisting of n distinct positive integers such that $\left(1-\frac{1}{s_1}\right)\left(1-\frac{1}{s_2}\right)\ldots\left(1-\frac{1}{s_n}\right)=\frac{51}{2010}$
Ground Truth: 39

Qwen3-30B-A3B-Thinking (29,074 tokens)	**Qwen3-30B-A3B-TIES** (21,702 tokens)
However, *exhaustive analysis shows that no combination of replacing elements in the set $\{1, 2, ..., 39\}$ results in the exact target value due to constraints on integer solutions. - Thus, $n = 39$ is not sufficient, and $n = 40$ is the smallest possible*	*We need this to be ≈1/1.0149≈0.9853.* So [k(m+1)]/[m(k+1)] = 0.9853 ⇒ k(m+1) = 0.9853m(k+1) ⇒ km + k = 0.9853km + 0.9853m ⇒ 0.0147km + k - 0.9853m = 0. For k=39: 0.0147*39m + 39 - 0.9853m = 0.5733m + 39 - 0.9853m = -0.412m + 39 = 0 ⇒ m=39/0.412≈94.66. *Try m=95: [39*96]/[95*40]=3744/3800=0.98526* , which is very close to 0.9853. . . .
Answer: 40 ×	**Answer:** Therefore, the minimal n is 40 ×
Qwen3-30B-A3B-MI-0.8 (29,974 tokens)	**Qwen3-30B-A3B-S^3** (8,513 tokens)
So 445/429 ¿ 335/323 ¿ 364/351. *Is there a k where (40/39)*(k+1)/k = 335/323?* Solve for k: (40/39)(k+1)/k = 335/323 ⇒ (k+1)/k = (335/323)*(39/40) = (335*39)/(323*40) = (5*67*3*13)/(17*19*8*5) = (67*3*13)/(17*19*8) = 2613/2584. So (k+1)/k = 2613/2584 ⇒ k = 2584/(2613-2584) = 2584/29 ≈89.103, not integer. *So closest integers are k=89 and k=90, but neither give the exact value.*	Oh! That's a key insight I missed earlier. **So the product of (1 - 1/k) for k=2 to m is 1/m.**
Answer: 39 ✓	**Answer:** 39 ✓

Table 13: **MATH-500, Example 4: Volume Enclosed by Three Distances** (41.7% token reduction). S^3 expands the sum once — the cross terms cancel and $x^2 + y^2 + z^2$ falls out; Thinking and MI derive the same sphere and then answer with its name instead of its volume.

Problem: For a point P, let d_1, d_2 and d_3 represents the distances from P to the planes $x - z = 0$, $x - 2y + z = 0$, and $x + y + z = 0$. Let S be the set of points P such that $d_1^2 + d_2^2 + d_3^2 = 36$. Find the region of the volume enclosed by S.
Ground Truth: 288π

Qwen3-4B-Thinking (16,803 tokens)
Wait, wait! *Wait a second—I just realized: in the problem statement, it says "the region of the volume enclosed by S". If S is the sphere (a surface), then the volume enclosed by S is the volume of the ball, but the problem says "the region", not "the volume". So they want the region (the set of points), which is the ball.* In standard math notation, the set of points with $x^2 + y^2 + z^2 \leq 36$ is the closed ball of radius 6 centered at the origin.

Answer: $\boxed{\text{solid sphere of radius 6}}$ ×

Qwen3-4B-TIES (8,188 tokens)
Let me verify that again because that seems surprisingly clean. *Let's check each expansion again to make sure I didn't make a mistake.* First term: $3(x - z)^2 = 3x^2 - 6xz + 3z^2$ Second term: $(x - 2y + z)^2 = x^2 + (-2y)^2 + z^2 + 2*(x*(-2y) + x*z + (-2y)*z) = x^2 + 4y^2 + z^2 - 4xy + 2xz - 4yz$ . . .

Answer: $\boxed{288\pi}$ ✓

Qwen3-4B-MI-0.8 (9,979 tokens)
Given that, I think the answer is the solid ball (closed ball) of radius 6 centered at the origin. *To write it neatly in a box, maybe using standard notation. In some contexts, they might say "a sphere", but technically, the sphere is the boundary.* Wait, but the problem says "the region of the volume enclosed by S", so S is the sphere (boundary), so enclosed volume is the ball.

Answer: $\boxed{x^2 + y^2 + z^2 \leq 36}$ ×

Qwen3-4B-S^3 (4,772 tokens)
Wow, all cross terms cancel out! That's nice. **So total = $6x^2 + 6y^2 + 6z^2$** . . .

Answer: $\boxed{288\pi}$ ✓

Table 14: **CMIMC25, Example 5: Trinomial Coefficients Modulo 3** (47.4% token reduction). S^3 substitutes $(1 + x + x^2) \equiv (x - 1)^2$ before applying Lucas; TIES and MI reduce the exponent by base-3 digits and then miscount the coefficients.

Problem: The binomial coefficient $\binom{n}{k}$ can be defined as the coefficient of x^k in the expansion of $(1 + x)^n$. Similarly, define the trinomial coefficient $\binom{n}{k}_3$ as the coefficient of x^k in the expansion of $(1 + x + x^2)^n$. Determine the number of integers k with $0 \le k \le 4048$ such that $\binom{2024}{k}_3 \equiv 1 \bmod 3$.
Ground Truth: 648

Qwen3-4B-Thinking (28,852 tokens) *Wait, is that right? Let's list the multipliers again:* i=0:1, i=1:3, i=2:3, i=3:3, i=4:1, i=5:1, i=6:3, i=7:1. *Multiply step by step: 1*3=3; 3*3=9; 9*3=27; 27*1=27; 27*1=27; 27*3=81; 81*1=81.* Yes, 81. *But wait, let's sanity check with smaller n to make sure our approach isn't flawed.* . . .	**Qwen3-4B-TIES** (24,500 tokens) Number of valid choices for a_m=2 is: - If there are s such digits, it's 2^{s-1} when s¿0, and 1 when s=0. *Wait for s=1: 2^0=1 s=2:2^1=2 s=0:1 Now let's test:* n=1: base3=1 → a_0=1, s=0 → count=3^1 *1=3 n=2: base3=2 → a_0=2, s=1 → count=3^0 *1=1 *(actual is 2)* . . .
Answer: 648 ✓	**Answer:** 32 ×
Qwen3-4B-MI-0.8 (31,669 tokens) *Reading remainders from bottom to top:* $2024_{10} = 2202222_3$ *So the base-3 digits are: [2, 2, 0, 2, 2, 2, 2]* Count: - Digits equal to 0: 1 - Digits equal to 1: 0 - Digits equal to 2: 6 . . .	**Qwen3-4B-S^3** (15,184 tokens) Let's verify: $(x-1)^2 = x^2 - 2x + 1 = x^2 + x + 1$ in $\mathbb{F}_3$, since $-2 \equiv 1 \mod 3$. **Yes! Perfect. So in $\mathbb{F}_3[x]$, $x^2 + x + 1 = (x-1)^2$. Therefore, $(1 + x + x^2)^n = (x-1)^{2n}$ in $\mathbb{F}_3[x]$.**
Answer: 64 ×	**Answer:** 648 ✓

Table 15: **AIME26, Example 6: Sets with Exactly 4040 Cousins** (51.2% token reduction). S^3 finds the multiplicative chain decomposition; Thinking and TIES reason from powers of two.

Problem: Call finite sets of integers S and T cousins if - S and T have the same number of elements, - S and T are disjoint, and - the elements of S can be paired with the elements of T so that the elements in each pair differ by exactly 1. For example, $\{1, 2, 5\}$ and $\{0, 3, 4\}$ are cousins. Suppose that the set S has exactly 4040 cousins. Find the least number of elements the set S can have.
Ground Truth: 107

Qwen3-4B-Thinking (31,320 tokens)	**Qwen3-4B-TIES** (24,345 tokens)
If we want 4040 cousins, which is 4040 = 2^3 * 5 * 101, but since the only possible cousin counts are powers of 2 (for good separation sets), the problem must have a typo, and the intended number is 4096 (2^{12}), so the minimal size is 12. *But wait, no—another idea: in the problem, when S has a block of size 2, it "uses up" two elements to get 1 cousin, whereas a single element uses one element to get 2 cousins.*	Even if we try to reduce $C(S)$ by introducing conflicts (e.g., bringing two elements to distance 2), we reduce $C(S)$ by 1 per conflict. - To get $C(S) = 4040$ from $2^{12} = 4096$, we'd need 56 such conflicts. - But each conflict requires two elements, so we'd need at least $2 \times 56 = 112$ elements $\rightarrow n \geq 112$ - This is much larger than 4039. *Thus, no smaller n can yield exactly 4040 cousins using any reasonable construction.*
Answer: 12 ×	**Answer:** 4039 ×
Qwen3-4B-MI-0.8 (31,687 tokens)	**Qwen3-4B-S^3** (15,456 tokens)
Wait, let's use the k=2 example: Block 1: left=-1, right=1 Block 2: left=1, right=3 Configuration 1: (-1,1) – block1 left, block2 left Configuration 2: (-1,3) – block1 left, block2 right Configuration 3: (1,3) – block1 right, block2 right So: L(1)=1 (block1 left), R(1)=1 (block1 right) L(2)=1 (block2 left: only in config 1), R(2)=2 (block2 right: configs 2,3) *Wait, no, L(2) is number of configs where block2 chooses left, which is 1; R(2)=2.* . . .	Yes! From the earlier examples: m=1: 2 = 1+1 m=2: 3 = 2+1 m=3: 4 = 3+1 And for a chain of length m, the valid assignments are sequences where once you choose right, all subsequent must choose right, so m+1 choices. Perfect! **Therefore, if S is partitioned into k disjoint chains (each chain being a maximal set of elements spaced by 2), then the total number of cousins is the product of (length of chain + 1) over all chains.**
Answer: 107 ✓	**Answer:** 107 ✓

Table 16: **Olympiad-Bench, Example 7: Maximising a Jump in a Moving Average** (36.0% token reduction). S^3 pins down when a term can equal 1, then takes the largest averaging window at the final step; the baselines take the smallest.

Problem: Let $a_0, a_1, a_2, \ldots$ be a sequence of real numbers such that $a_0 = 0, a_1 = 1$, and for every $n2$ there exists $1kn$ satisfying $a_n = \frac{a_{n-1}+\cdots+a_{n-k}}{k}$ Find the maximal possible value of $a_{2018} - a_{2017}$.
Ground Truth: $\frac{2016}{2017^2}$

Qwen3-30B-A3B-Thinking (20,465 tokens)
Similarly, in Path B, to get d4¿0, we had a3 ¡ a1 + a2? *Wait, no, earlier for d4¿0 with k=3 at n=4, we had a4=(a1+a2+a3)/3 ¿a3 ⇒ a1+a2 ¿2a3.* To maximize d4=(a4 -a3)= (a1+a2 -2a3)/3, need to minimize (a1+a2 -2a3), i.e., minimize a3 (since a1,a2 fixed once chosen). a1=1, a2 was minimized to 1/2 at n=2, then at n=3, to minimize a3, we take k=1 (a3=a2=1/2), so a3=1/2, hence d4=(1 + 1/2 - 2*(1/2))/3=(1 + 1/2 -1)/3=(1/2)/3=1/6, which is exactly what we got. *And this is the maximum possible d4 because we minimized a3 (the variable term).*

Answer: $\boxed{\frac{1}{4034}}$ ×

Qwen3-30B-A3B-TIES (31,161 tokens)
For even n, $a_n = 2/3(1 - 1/2^n) < 2/3$ So $a_2 = 2/3(1 - 1/4) = 1/2 = 0.5$ $a_4 = 2/3(1 - 1/16) = 15/24 = 0.625$ $a_6 = 2/3(1 - 1/64) = 63/96 = 0.65625$ So it's increasing towards 2/3. *So the even-indexed terms are increasing: $a_2 = 0.5$, $a_4 = 0.625$, $a_6 = 0.65625$, ..., approaching 2/3 ≈ 0.666...*

Answer: 0 ×

Qwen3-30B-A3B-MI-0.8 (29,032 tokens)
The best strategy to achieve a large difference at position n is: - Keep the sequence constant at 1 from a_1 up to a_{n-2}, - At step $n-1$, compute $a_{n-1} = \frac{\text{sum of } a_0 \text{ to } a_{n-2}}{n-1}$, which gives $a_{n-1} = \frac{n-2}{n-1}$, *- Then at step n, take $k = 2$:* $a_n = \frac{a_{n-1}+a_{n-2}}{2} = \frac{(n-2)/(n-1)+1}{2} = \frac{2n-1}{2(n-1)} \cdots$

Answer: $\boxed{\frac{1}{4034}}$ ×

Qwen3-30B-A3B-S^3 (13,098 tokens)
Moreover, **the only way to get a_n=1 is if all the terms being averaged (the last k terms) are equal to 1, because if any term in the average is less than 1, the average would be less than 1.**

Answer: $\boxed{Sothemaximalpossiblevalueis\ldots}$ ✓

Table 17: **HMMT25, Example 8: A Polynomial Coefficient by Interpolation** (18.8% token reduction). S^3 exploits $P'(x) \equiv -1$ in the interpolation; the baselines replace the coefficient with a count of the non-zero values.

Problem: Let f be the unique polynomial of degree at most 2026 such that for all $n \in \{1, 2, 3, \ldots, 2027\}$, $f(n) = \begin{cases} 1 & \text{if } n \text{ is a perfect square} \\ 0 & \text{otherwise} \end{cases}$ Suppose that $\frac{a}{b}$ is the coefficient of x^{2025} in f, where a and b are integers such that $\gcd(a, b) = 1$. Compute the unique integer r between 0 and 2026 (inclusive) such that $a - rb$ is divisible by 2027. (Note that 2027 is prime.)
Ground Truth: 1037

Qwen3-4B-Thinking (18,715 tokens) *The leading coefficient C of $f(x)$ satisfies: $C \equiv -m \mod p$* where m is the number of perfect squares in $\{1, 2, \ldots, p\}$ *This result is verified through multiple test cases with small primes and confirmed via detailed analysis* . . .	**Qwen3-4B-TIES** (22,348 tokens) *So $f(x)$ is degree at most 44? Wait, hold on! This is a critical point.* The problem says "degree at most 2026", but if we have only 45 points where it's 1 and others 0, *isn't the interpolating polynomial of degree at most 44?* . . .
Answer: 1982 ×	**Answer:** 0 ×
Qwen3-4B-MI-0.8 (13,991 tokens) Therefore: $c_{2025} \equiv -45 \mod 2027$. *To get the representative in $[0, 2026]$, add 2027: $r = (-45) \mod 2027 = 2027 - 45 = 1982$.* — Final Answer . . .	**Qwen3-4B-S^3** (11,361 tokens) Differentiating both sides: $P'(x) = 2027x^{2026} - 1 \equiv -1 \mod 2027$ **So $P'(n) = -1$ for all $n \in \mathbb{F}_{2027}$.**
Answer: 1982 ×	**Answer:** 1037 ✓

Table 18: **Olympiad-Bench, Example 9: Sum of Squares over the Rationals** (53.6% token reduction). S^3 reduces the question to how many rational squares 7 needs and answers with Legendre's theorem; Thinking and TIES try to exhibit a construction instead.

Problem: Find the smallest number n such that there exist polynomials $f_1, f_2, \ldots, f_n$ with rational coefficients satisfying $x^2 + 7 = f_1(x)^2 + f_2(x)^2 + \cdots + f_n(x)^2$.
Ground Truth: 5

Qwen3-30B-A3B-Thinking (31,456 tokens)
I think I have to trust the theory here. *The key points are: - n=1: impossible, as shown. - n=2: impossible, as shown (requires 7 to be sum of two rational squares, which it's not). - n=3: impossible, as shown (7 cannot be sum of three rational squares). - n=4: possible, because 7 can be written as sum of four rational squares, and the orthogonality condition is solvable over Q (local conditions satisfied, so global solution exists by Hasse-Minkowski).* Therefore, the minimal n is 4.

Answer: 4 ×

Qwen3-30B-A3B-TIES (15,776 tokens)
Yes! *Here's a known example:* $x^2 + 7 = \left(\frac{x}{2} + \frac{3}{2}\right)^2 + \left(\frac{x}{2} + \frac{3}{2}\right)^2 + \left(\frac{x}{2} - \frac{3}{2}\right)^2 + \left(\frac{x}{2} - \frac{1}{2}\right)^2$. *Let's check: Each term: -* $\left(\frac{x}{2} + \frac{3}{2}\right)^2 = \frac{x^2}{4} + \frac{3x}{2} + \frac{9}{4}$, *- Same for the second, -* $\left(\frac{x}{2} - \frac{3}{2}\right)^2 = \frac{x^2}{4} - \frac{3x}{2} + \frac{9}{4}$, *-* $\left(\frac{x}{2} - \frac{1}{2}\right)^2 = \frac{x^2}{4} - \frac{x}{2} + \frac{1}{4}$. *Sum: -* x^2*:* $4 \times \frac{1}{4} = 1$, *-* x*:* $\frac{3}{2} + \frac{3}{2} - \frac{3}{2} - \frac{1}{2} = (3-2)/2 = 1/2$, *- Constants:* $2 \times \frac{9}{4} + \frac{9}{4} + \frac{1}{4} = \frac{18+9+1}{4} = 28/4 = 7$. *Still has* x *term. But if we swap signs in the last two: Try: -* $f_1 = \frac{x}{2} + \frac{3}{2}$, *-* $f_2 = \frac{x}{2} + \frac{3}{2}$, *-* $f_3 = -\frac{x}{2} + \frac{3}{2}$, *-* $f_4 = -\frac{x}{2} + \frac{1}{2}$. *Then* x *terms:* $\frac{3}{2} + \frac{3}{2} + \frac{3}{2} + \frac{1}{2} = 4$. Worse. . . .

Answer: 4 ×

Qwen3-30B-A3B-MI-0.8 (32,175 tokens)
From the second equation: $ab = -cd \Rightarrow b = -\frac{c}{a}d$ (assuming $a \neq 0$). *Substitute into the third equation:* $b^2 + d^2 = \left(\frac{c^2}{a^2} + 1\right) d^2 = \frac{a^2+c^2}{a^2} d^2 = \frac{1}{a^2} d^2$. *So:* $\frac{1}{a^2} d^2 + e^2 + f^2 = 7$. *Let* $u = \frac{d}{a}$, *then* $u^2 + e^2 + f^2 = 7$. *This implies that 7 must be written as a sum of three rational squares, which is impossible (as shown earlier). Therefore, this configuration fails.* Trying other configurations (e.g., three linear, one constant) similarly leads to contradictions or irrational coefficients . . .

Answer: 5 ✓

Qwen3-30B-A3B-S^3 (14,931 tokens)
But Legendre's three-square theorem says a natural number n is expressible as sum of three integer squares iff n is not of the form $4^a(8b + 7)$. **7 is exactly of the form 8b + 7 (b=0), so 7 cannot be written as sum of three integer squares.** But can it be written as sum of four integer squares? **Yes, by Lagrange's four-square theorem, every natural number is sum of four integer squares.** In fact, $7 = 2^2 + 1^2 + 1^2 + 1^2 = 4 + 1 + 1 + 1 = 7$.

Answer: 5 ✓

C.2 VISION LANGUAGE REASONING

These six examples all turn on a figure — a food web, a clock face, a composite scene, a geometric diagram, and two accounting tables — and in four of the six no baseline answers correctly at all; in the other two exactly one does. The three analysed below are Table 20, where the hour hand is read as pointing at the wrong numeral; Table 22, where a spatial relation is never resolved; and Table 23, where the baselines lose track of which quantity the table leaves blank.

Representative examples.

Example 1 (Table 20): Reading an Analogue Clock (81.5% reduction, incorrect → correct). A clock face, four candidate times, and one thing to read: where the two hands point. The hour hand sits between the 4 and the 5, and the minute hand on the 6, so the time is 4:30 and the answer is (C) — which is what S^3 reports, in 394 tokens. The three baselines all report 5:30, and all three arrive there the same way: they read the hour hand as pointing *at* the 5 rather than between the 4 and the 5.

The traces are worth quoting because the error is visible in the first sentence and never corrected. MI-0.8 (4,041) opens *"It's pointing between 4 and 5. Wait, no, actually, the hour hand is pointing towards the 5. Wait, no, wait."* — the correct reading is produced and immediately discarded, and it is not recovered. TIES (2,130) reaches *"So answer D. But wait, maybe the hour hand is pointing to 5, which is 5:00, but the minute hand is at 6, so 30 minutes. So 5:30. Yes, that makes sense. So the answer is D."* and then re-derives the same conclusion twice more. Thinking (8,073) resolves the ambiguity the other way — *"The hour hand is positioned between 5 and 6"*, which the image does not support — and closes by explaining why a hand at 5:30 should sit half way between the 5 and the 6: a rule stated correctly and applied to the wrong pair of numerals.

S^3 reads the hands once and keeps what it read: *"the hour hand is between 4 and 5, the time must be 4:30"*. It then double-checks the one thing that could sink the reading — *"sometimes people get confused if the hour hand is exactly on a number. But here, the hour hand is just past 4"* — and stops. The 81.5% is not a shorter inspection; all four traces inspect the same pixels. It is the cost of the other three reopening a reading.

Example 2 (Table 22): Tandem Bikes and Matte Trucks (62.7% reduction, incorrect → correct). A comparative question about a cluttered scene: are there fewer tandem bikes behind the brown metal bicycle than matte trucks to the left of the green object? The answer is no, because both counts are 1 — the purple tandem bike is behind the brown bicycle, and the gray pickup is the matte truck to the left of the green truck. S^3 counts 1 and 1 and reports (B) in 551 tokens. The three baselines all count 0 and 1, all report (A) Yes, and all three make the same mistake in the same place.

That place is a single word. MI-0.8 (2,124) decides *"The purple tandem bike is to the right of the brown bicycle, so it's in front of the brown bicycle"* — it reads a left/right position off the image and converts it directly into a front/behind relation, which the image does not license. Thinking (7,059) reaches the same count through a longer route that also fixes the wrong colour for the brown bicycle before correcting itself, and TIES (1,478) arrives there through a chain of hedges: *"Wait, no. The large gray truck is in the center, and the green truck is behind it? Wait, no. The green truck is a small one, so the large gray truck is to its left."* In all three traces the second count is right and the first is zero, so the comparison $0 < 1$ then decides the answer.

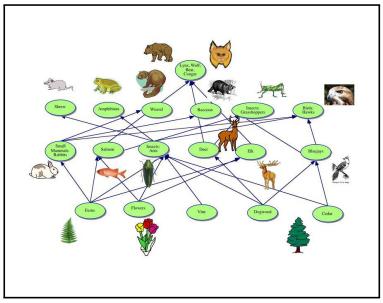
S^3 resolves the one relation the question depends on and then does the comparison: *"Behind it, there's a tandem bike"*, followed by *"number of tandem bikes behind brown bicycle: 1"*. It still has to decide whether a gray pickup counts as matte and whether the bus counts as a truck, and it gets both of those right too — but those two decisions are the same in all four traces. The 62.7% is the cost of the front/behind relation not being settled.

Example 3 (Table 23): A Missing Figure in a Trial Balance (75.6% reduction, incorrect → correct). An accounting table lists several companies, one cell of company D's row is blank, and four candidate amounts are offered. The blank cell is company D's *expenses*; company D's revenues, $1{,}054{,}116$, are printed in the table. So the problem is one substitution into net income = revenues − expenses + gains − losses, and S^3 performs exactly that, in 1,463 tokens, to get $1{,}023{,}309 + 58{,}275 = 1{,}081{,}584$ and option (C). All three baselines write down the same equation and then lose track of which of its terms is the one being asked for.

Thinking (10,382 tokens) works the equation for each option in turn, computing the expenses that would correspond to each candidate revenue, and finishes by choosing on magnitude: *"The options are all around 1 million, so 1,210,732 is the highest option. Wait, maybe the correct answer is D."* TIES (6,003) computes the same four candidate expenses and then asks itself *"Now, is there any other information?"* before settling on A. MI-0.8 (8,611) is the most explicit about the failure — it cannot decide whether the missing cell is expenses or revenues, says so, and then answers anyway: *"maybe the question is to find Expenses, but the options are Revenues. No. Wait, maybe the question is to find the missing amount for Company D, which is Expenses, and the correct answer is A."* Three traces, three different wrong options, and in each the arithmetic is sound.

S^3 spends its first line identifying which cell is blank and then never revisits the question: it reads company D's revenues off the table, substitutes once, and matches the result to the option list. The 75.6% is the difference between one substitution and four.

Table 19: **MathVista-testmini, Example 10: Ferns and Salmon in a Food Web** (87.1% token reduction). S^3 follows the one edge the question turns on; the baselines reconstruct the whole food web from memory.



Problem: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: In the diagram of the food web shown, if the number of ferns decrease, the supply of salmon will most likely? Choices: (A) decrease (B) can't tell (C) stay same (D) increase

Ground Truth: decrease

Qwen3-VL-4B-Thinking (5,741 tokens)
Wait, maybe the user made a mistake in the diagram description, but assuming the diagram is correct, we have to go with the connections given. Since the diagram shows Ferns connected to Rabbits, and Salmon connected to Insects: Ants (connected to Flowers), there's no direct or indirect connection between Ferns and Salmon. Therefore, the supply of Salmon would stay the same. So the answer is C) stay same.

Answer: C ×

Qwen3-VL-4B-TIES (2,167 tokens)
Therefore, we can't tell. Wait, but let's check again. The diagram has arrows from Ferns to Small Mammals: Rabbits. So rabbits eat ferns. Then, what do rabbits eat? If ferns are the main food, then fewer ferns → fewer rabbits. *Now, do rabbits eat salmon?* . . .

Answer: B ×

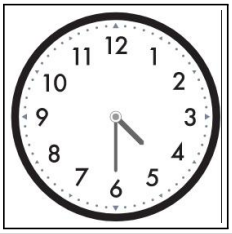
Qwen3-VL-4B-MI-0.8 (2,705 tokens)
Is there an arrow from Ferns to Salmon? Let me think. If the number of ferns decreases, what happens to salmon? *If ferns are a food source for salmon, then less ferns would mean less food for salmon, so salmon population would decrease.*

Answer: C ×

Qwen3-VL-4B-S^3 (280 tokens)
Wait, in the diagram, there's a line from ferns to salmon. So ferns are a food source for salmon. If ferns decrease, that means there's less food available for salmon. **So salmon's supply would likely decrease because they have less food.**

Answer: A ✓

Table 20: **MathVista-testmini, Example 11: Reading an Analogue Clock** (81.5% token reduction). S^3 reads both hands once; the baselines re-open the hour hand a dozen times.



Problem: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: What time does the clock show? Choices: (A) 9:30 (B) 1:30 (C) 4:30 (D) 5:30 (E) 11:30

Ground Truth: 4:30

Qwen3-VL-4B-Thinking (8,073 tokens)
Wait, this is really hard without seeing the image. Wait, but the user provided a description: "the clock show" with numbers 1-12. The hour hand is pointing to 5. Or wait, no. . . .

Answer: D ×

Qwen3-VL-4B-TIES (2,130 tokens)
This is a bit tricky. Wait, perhaps the image shows the hour hand pointing to 5. But if the minute hand is at 6 (30 minutes), then the hour hand should be halfway between the current hour and the next. So if it's 5:30, . . .

Answer: D ×

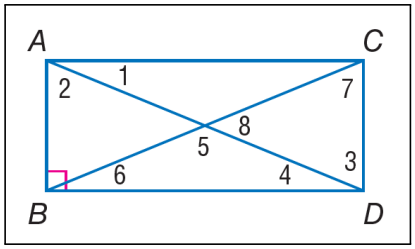
Qwen3-VL-4B-MI-0.8 (4,041 tokens)
Wait, the user provided a clock image. Let me try to imagine it. The clock has numbers 1 through 12. The minute hand is pointing at the 6 (which is 30 minutes), so the time is X:30.

Answer: D ×

Qwen3-VL-4B-S^3 (394 tokens)
Looking at the image, **the hour hand is pointing just past the 4, and the minute hand is pointing at the 6.**

Answer: C ✓

Table 21: **MathVista-testmini, Example 12: Which Angle Is Numbered Two** (67.3% token reduction). S^3 takes the two labels at A as the two parts of the right angle and subtracts once; Thinking and MI argue over the vertex order in "ABDC" and set angle 2 equal to angle 1.



Problem: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Quadrilateral $ABDC$ is a rectangle. If $m\angle 1 = 38$, find $m\angle 2$ Choices: (A) 33 (B) 38 (C) 52 (D) 87

Ground Truth: 52

Qwen3-VL-2B-Thinking (16,634 tokens)
The problem says quadrilateral ABDC is a rectangle, so ABDC is a rectangle, which means AB is parallel to CD, and AD is parallel to BC. Also, all angles in a rectangle are right angles, so angle at B is 90 degrees, right? *Wait, the diagram shows a rectangle with points A, B, D, C, so the vertices are A, B, D, C, making ABDC a rectangle. So AB is one side, BD is a diagonal, and CD is the opposite side. Wait, no, in a rectangle, the order of the vertices matters.* ...

Answer: B ×

Qwen3-VL-2B-TIES (7,072 tokens)
First, let's recall what we know about rectangles. A rectangle has all right angles, so angle at B is 90 degrees. Also, in a rectangle, opposite sides are equal, and diagonals are equal and bisect each other. *Wait, but here we have quadrilateral ABDC as a rectangle. Wait, let's check the diagram. First, let's identify the rectangle. The diagram shows a rectangle with vertices A, B, D, C. Wait, the vertices are A, B, D, C. So ABDC is a rectangle.* ...

Answer: C ✓

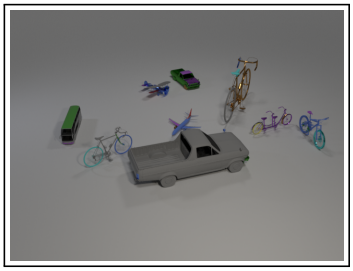
Qwen3-VL-2B-MI-0.8 (18,735 tokens)
Wait, hold on, the vertices are A, B, D, C? Wait, the diagram shows a rectangle with points A, B, C, D. Let me check the labels. The rectangle is ABDC? Wait, no, the rectangle is probably ABCD, but the problem says ABDC. Wait, maybe it's a typo, but let's look at the diagram. The diagram has points A, B, C, D with a right angle at B, so AB is a side, BC is a side, CD is the other side, DA is the last side. Wait, the right angle is at B, so angle ABC is 90 degrees. That makes sense for a rectangle: all angles are 90 degrees, opposite sides equal. ...

Answer: B ×

Qwen3-VL-2B-S^3 (2,315 tokens)
Now consider $\angle 2$, which is at vertex A, between side AB and diagonal AD. Since AB and AD are adjacent sides of the rectangle, $\angle BAD = 90°$. We already have: $\angle BAC = 38°$ So: $\angle CAD = \angle BAD - \angle BAC = 90° - 38° = 52°$ Thus: $\angle 2 = \angle CAD = 52°$ ...

Answer: C ✓

Table 22: **MathVista-testmini, Example 13: Tandem Bikes and Matte Trucks** (62.7% token reduction). S^3 resolves "behind" once and compares; the baselines never settle the spatial relation.



Problem: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Are there fewer tandem bikes that are behind the brown metal bicycle than matte trucks on the left side of the green object? Choices: (A) Yes (B) No

Ground Truth: No

Qwen3-VL-2B-Thinking (7,059 tokens)
Let's list all objects: 1. Small green bus (left side, gray and green) 2. Gray road bike (left of the truck) 3. Blue airplane (middle) 4. Small blue and green truck (green object? Wait, the green object is the truck with green color. Wait, the green truck is the one with the green color, so the truck is green. Then, the other truck is the gray pickup truck in the center. . . .

Answer: A ×

Qwen3-VL-2B-TIES (1,478 tokens)
Wait, "behind" in 3D space—relative to the brown bike. The brown bike is a road bike, and the tandem bike is a little to the right and maybe behind? Wait, no. Let's visualize. . . .

Answer: A ×

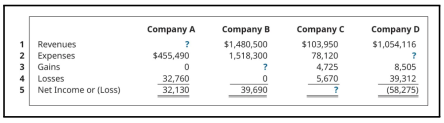
Qwen3-VL-2B-MI-0.8 (2,124 tokens)
so behind the brown bicycle? Wait, no—position matters. The brown bicycle (road bike) is on the right side. *The tandem bike is to the right of the brown bicycle, so maybe behind? Wait, no. Let's look at the scene.* . . .

Answer: A ×

Qwen3-VL-2B-S^3 (551 tokens)
Let's see: **the tandem bike is purple and located behind the brown bicycle. So that's one tandem bike.**

Answer: B ✓

Table 23: **MMMU, Example 14: A Missing Figure in a Trial Balance** (75.6% token reduction). S^3 identifies which cell is missing; the baselines solve for a cell that was already given.



	Company A	Company B	Company C	Company D
1 Revenues	?	$1,480,500	$103,950	$1,054,116
2 Expenses	$455,490	1,518,300	78,120	?
3 Gains	0	?	4,725	8,505
4 Losses	32,760	0	5,670	39,312
5 Net Income or (Loss)	32,130	39,690	?	(58,275)

Problem: Each of the following situations relates to a different company. [Image 1] For company D, find the missing amounts. A. $1,034,325B$.1,098,650 C. $1,081,584D$.1,210,732 Answer with the option letter only.

Ground Truth: C

Qwen3-VL-4B-Thinking (10,382 tokens)
Wait, maybe the question is asking for Revenues? Let's see. Wait, let's rearrange the formula for Revenues: Revenues = Net Income + Expenses - Gains + Losses Wait, let's see. From Net Income = Revenues - Expenses + . . .

Answer: D ×

Qwen3-VL-4B-TIES (6,003 tokens)
Since all options satisfy the equation, but the problem must have one correct answer, maybe I made a mistake in the problem . . .

Answer: A ×

Qwen3-VL-4B-MI-0.8 (8,611 tokens)
So all options work. But that can't be. There must be a mistake in the problem or my reasoning. Wait, maybe the question is to find the missing amount for Company D, which is Expenses, and the . . .

Answer: A ×

Qwen3-VL-4B-S^3 (1,463 tokens)
Wait, so the only missing amount for Company D is Expenses. Let's verify. Net Income = Revenues - Expenses + Gains - Losses **We can rearrange to solve for Expenses: Expenses = Revenues + Gains - Losses**

Answer: C ✓

Table 24: **MMMU, Example 15: Shifts in a Loanable-Funds Diagram** (95.5% token reduction). S^3 moves one curve per shock and reads the new crossing off the diagram; MI sends the fall in capital spending the wrong way and TIES moves the rise in saving onto the demand curve.

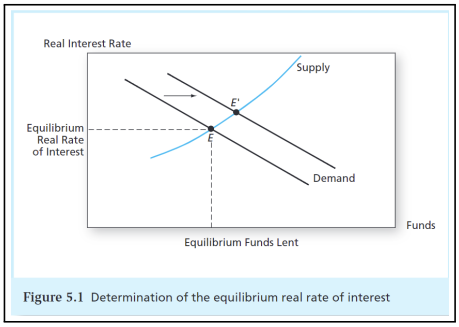


Figure 5.1 Determination of the equilibrium real rate of interest

Problem: Use the image to analyze the effect of the following on the level of real interest rates:[Image 1] a. Businesses become more pessimistic about future demand for their products and decide to reduce their capital spending. b. Households are induced to save more because of increased uncertainty about their future Social Security benefits. c. The Federal Reserve Board undertakes open-market purchases of U.S. Treasury securities in order to increase the supply of money. A. a.increase ;b.decrease ;c.increase B. a.increase ;b.decrease ;c.decrease C. a.decrease ;b.decrease ;c.decrease D. a.decrease ;b.increase ;c.decrease Answer with the option letter only.

Ground Truth: C

Qwen3-VL-2B-Thinking (7,620 tokens)
So for option a: demand for funds (for investment) decreases → demand curve shifts left → intersection is to the left of E → interest rate is lower (decrease). *Option b: households save more → demand for funds (for investment) decreases → demand curve shifts left → interest rate is lower (decrease).* Option c: Federal Reserve buys Treasury securities → supply of money increases → supply curve shifts right → intersection is to the right of E → interest rate is lower (decrease). So all three options decrease the interest rate. . . .

Answer: C ✓

Qwen3-VL-2B-TIES (919 tokens)
D. a.decrease ;b.increase ;c.decrease → yes, that's D. Wait, let me confirm again. For a: businesses reduce capital spending → demand for funds decreases → demand curve shifts left → equilibrium real rate decreases → correct. *For b: households save more → demand for funds increases → demand curve shifts right → equilibrium real rate increases → correct.* For c: Fed buys Treasury securities → increases money supply → supply of funds shifts right → equilibrium real rate decreases → correct. So the answer is D. . . .

Answer: D ×

Qwen3-VL-2B-MI-0.8 (1,023 tokens)
In the supply and demand model for funds, the demand curve is the demand for funds by businesses, households, etc. *If businesses reduce their capital spending, that's a decrease in the demand for funds (since they're not borrowing as much to spend on capital). So the demand curve shifts left. The intersection of supply and demand shifts left, so the equilibrium real rate of interest increases. Wait, but in the graph, the supply curve is upward sloping, and demand is downward sloping.* . . .

Answer: B ×

Qwen3-VL-2B-S^3 (341 tokens)
So a is decrease. b. Households save more due to uncertainty about Social Security benefits. **Saving more means more funds in the banking system, so the supply of funds increases. Supply of funds shifts right.** That would lead to a lower real interest rate. Because with more supply, the equilibrium interest rate falls. So b is decrease. c. Federal Reserve buys Treasury securities, increasing money supply. **That shifts the supply curve for funds to the right (since more money is in the economy). So supply increases, equilibrium interest rate decreases.** So c is decrease. . . .

Answer: C ✓

C.3 AUDIO REASONING

These five examples are audio questions: one listening-comprehension item, three from the MMAR set of audio-reasoning questions, and one disfluency-annotation item from MMSU. In four of the five no baseline answers correctly; the fifth, Table 27, is a clip on which all four methods succeed, and it is included for contrast. Each of the other four turns on a single cue, and each of the three analysed below turns on a different kind: Table 26, where the culprit is identified by what is said rather than by how someone reacts; Table 28, where the baselines discount a demonstration that has already been given; and Table 29, where one opening sound has to be classified before anything else can be counted.

Representative examples.

Example 1 (Table 26): Who Ate the Powdered Donut (55.0% reduction, incorrect → correct). A four-speaker exchange, and one question: who ate the donut? The fourth speaker says so — *"Yeah, that was me."* — and S^3 quotes exactly that, naming the fourth person in 218 tokens. All three baselines hear the same exchange, all three reconstruct the culprit from something other than the admission, and all three land on the wrong speaker.

Thinking (3,929 tokens) and MI-0.8 (3,006) build their answer on a cough. MI-0.8 reasons *"The physical reaction (coughing) aligns with the act of consuming something like a powdered donut, and Speaker 3's denial followed by this sound indicates guilt"*; Thinking lays the same inference out as a four-row elimination and concludes *"the third speaker (Speaker 3) ate the powdered donut"*. The cough is real, and the inference from it is not unreasonable — but it is a guess about a speaker who never says anything, made in a clip where another speaker does. TIES (484) takes the third route, from ownership to guilt: *"The speaker who then makes the statement about it being 'his last donut' and calls the others 'disrespectful' is also the first speaker"*, and answers A.

S^3 transcribes the same four turns and reads the fourth one for what it is, an admission: *"Speaker 4 (fourth person) admits by saying, "Yeah, that was me.", "* It then confirms the fourth person is option D and stops. The 55.0% is the cost of the two competing inferences that S^3 does not enter.

Example 2 (Table 28): Counting the Chinese Tones Demonstrated (70.9% reduction, incorrect → correct). The speaker teaches four Mandarin tones, demonstrates each one individually, and then says six syllables that use them. The question asks how many different tones are *involved in the six pronunciation words*, and the answer key says four — which requires counting the individual demonstration, since the six words on their own use three. S^3 takes that reading and answers (D) in 535 tokens. All three baselines take the other reading and answer (A) 3.

What makes this example useful is that the baselines find the missing fourth tone and then discard it. MI-0.8 (5,912) states the whole case and then rules it out in one sentence: *"Although the speaker mentioned learning four tones earlier in the audio with the "a" sounds, the specific six words provided do not include the fourth tone."* TIES (1,841) and Thinking (4,147) reach the same conclusion without ever considering the demonstration, both listing the same three tones and both ending on the same negative finding — Thinking's is *"The fourth tone (à) is not demonstrated in any of the six syllables."* Each of the three then reports the count of the narrower set.

S^3 keeps both parts of the clip in scope: it records the four individual demonstrations and then the six words, and finds the fourth tone among the former. The 70.9% is the cost of the six-syllable inventory the baselines build.

Example 3 (Table 29): Types of Disfluency in the Audio (41.8% reduction, incorrect → correct). The clip opens with a short laugh, then a woman describes a trifle: *"Ha, well my sister is the simstrous of the two of us."* The options ask which disfluency categories are present, and the answer is A, discourse markers and restarts — *well* for the first, *simstrous* for the second. S^3 answers A in 271 tokens. All three baselines answer D, and all three do so because they classify the opening laugh as a filled pause.

The pair of traces to compare is S^3's and MI-0.8's, because they transcribe the same sound differently. MI-0.8 (1,944) hears *"Huh"* and files it with *um* and *uh*; TIES (1,459) does the same; Thinking (466) calls it *"a short, sharp laugh"* and then files it as a filled pause anyway — *"This is a common filled pause, often used to indicate"*. S^3 makes the same observation and draws the opposite conclusion: *"The initial "Ha" is a chuckle, which is a filler but not a standard filled pause*

like 'uh' or 'um'." The rest of the classification follows. MI-0.8 also shows the baselines overriding their own taxonomy — *"So even though discourse markers aren't disfluencies, the test might be considering them as such"* — to keep the option that the misclassification produced.

S^3 then checks the remaining options off and stops. The 41.8% is the smallest reduction of the twenty, and the difference between the two traces is a single perceptual decision taken in the first line of both.

Table 25: **MMAR, Example 16: A Profession Inferred from a Conversation** (42.4% token reduction). S^3 takes the word "test" as the domain cue; the baselines supply a fitness practice that does not exist.

Problem: What is the most likely profession of the first speaker? **Audio transcript** *(reader aid; the model received only the audio)*: "all right bro looking good we're done yeah we're all done man what's up there you go this is for you you want me to eat this yeah man candy yeah sure all right that was a test dude which you failed" **Ground Truth:** Dentist	
Qwen3-Omni-30B-A3B-Thinking (1,978 tokens) The key part is that they're talking about candy, and then one says it was a test that the other failed. But what does this have to do with professions? Now, the question is: What is the most likely profession of the first speaker? Options are Dentist, Teacher, Fitness Coach, Nurse. I need to think about contexts where someone might give candy as a test or something related to health. Let me consider each option: *- Dentist: Dentists often talk about oral health, and candy might be related to teeth. But dentists usually advise against eating candy because it's bad for teeth.* However, in this case, the first speaker gives candy and says it was a test. *Maybe it's a test for something like how people react to candy, but that seems odd.*	**Qwen3-Omni-30B-A3B-TIES** (549 tokens) *The first speaker says, 'All right, bro. Looking good. We're done?' The second speaker asks, 'We're done?' and the first speaker confirms, 'Yeah, we're all done, man.' The second speaker then says, 'What's up?' and the first speaker responds, 'There you go. This is for you.' The second speaker asks, 'You want me to eat this?' and the first speaker says, 'Yeah, man. Why not? The candy.* . . .
Answer: C ×	**Answer:** C ×
Qwen3-Omni-30B-A3B-MI-0.8 (7,446 tokens) *Perhaps for a dentist, after a check-up, they might say "Looking good," and then give a small treat, but candy is bad for teeth, so it doesn't make sense for a test.* Unless it's to see if the patient resists, but the patient is given candy to eat, so if they eat it, they fail, but dentists don't usually do that. For a teacher, it's unlikely. For a nurse, similar. Let's think about the word "test." *In medical contexts, there are glucose tests, but candy isn't used for that.*	**Qwen3-Omni-30B-A3B-S^3** (316 tokens) **Crucially, the word 'test' is used in what appears to be a professional context, suggesting a formal evaluation.**
Answer: C ×	**Answer:** A ✓

Table 26: **MMAR, Example 17: Who Ate the Powdered Donut** (55.0% token reduction). S^3 reads the fourth speaker's admission; the baselines infer the culprit from a cough.

Problem: Which speaker ate the powdered donut
Audio transcript *(reader aid; the model received only the audio)*: "Hey, wait my last powder donut it wasn't me wasn't me Guys that was his last powder donut. Don't be so disrespectful. Okay?"
Ground Truth: Fourth person

Qwen3-Omni-30B-A3B-Thinking (3,929 tokens)
Here, "his" refers to Speaker 1 (the owner), confirming the donut was eaten, and the phrase "Don't be so disrespectful" implies Speaker 4 is addressing the eater. *The coughing directly after Speaker 3's denial suggests they are the one who ate the donut, as the physical reaction gives away their guilt.* Speaker 1 is the owner and not the eater, Speaker 2 denied responsibility without coughing, and Speaker 4 is accusing others, so they are not the eater. Therefore, the third speaker (Speaker 3) ate the powdered donut.

Answer: B ×

Qwen3-Omni-30B-A3B-TIES (484 tokens)
Speaker 2 (female) says, 'It wasn't me,' which is a direct denial of eating it. Speaker 1 then says, 'Wasn't me,' which is another denial. *After these denials, Speaker 1 starts coughing, and then says, 'Guys, that was his last powdered donut. Don't be so disrespectful. Okay, fresh up right now.'* This is the key part. . . .

Answer: A ×

Qwen3-Omni-30B-A3B-MI-0.8 (3,006 tokens)
The audio: "Wasn't me." then "cough cough cough Oh!" then "Guys, that was his last powdered donut..." So coughing is separate from Speaker 4's speech. Therefore, the coughing is from one of the previous speakers. *Since Speaker 3 just said "Wasn't me," and then coughs, it's likely Speaker 3. Speaker 2 spoke earlier, so less likely. Therefore, Speaker 3 ate it.*

Answer: B ×

Qwen3-Omni-30B-A3B-S^3 (218 tokens)
Then, analyzing the audio reveals four speakers: Speaker 1 (first person) asks, "Hey, when the last powdered donut?", Speaker 2 (second person) denies, "It wasn't me.", Speaker 3 (third person) denies, "Wasn't me.", **and Speaker 4 (fourth person) admits by saying, "Yeah, that was me.", followed by coughing and gagging, then clarifying, "That was his last powdered donut. Don't be so disrespectful." This admission occurs when the speaker identifies themselves as the one who ate it.**

Answer: D ✓

Table 27: **MMAU_testmini, Example 18: The Ingredients of a Trifle** (59.6% token reduction). S^3 recalls the layers once and dismisses each wrong list in a clause; the baselines re-check every distractor ingredient by ingredient.

Problem: What ingredients did Speaker 1 say she used in the trifle?
Audio transcript *(reader aid; the model received only the audio)*: "it's a trifle it's got all of these layers it's got all of these layers first there's a layer of ladyfingers then a layer of jam then custard which i made from scratch then raspberries more ladyfingers then beef sauteed with peas and onions and then bananas and then i just put some whipped cream on top it tastes like feet i like it are you kidding"
Ground Truth: (D) Ladyfingers, jam, custard, raspberries, beef, peas, onions, bananas, whipped cream

Qwen3-Omni-30B-A3B-Thinking (1,373 tokens)
A. Bread, jelly, pudding, strawberries, chicken, carrots, garlic, apples, cream - *Bread? Not mentioned; she said lady fingers, which are like cookies or biscuits, but not bread. - Jelly? She said jam, which is similar, but option says jelly. - Pudding? She said custard, which might be confused, but custard is different. - Strawberries? She said raspberries. - Chicken? She said beef.* - Carrots, garlic, apples? Not mentioned. - Cream? She said whipped cream, but option says cream, which might be close. . . .

Answer: D ✓

Qwen3-Omni-30B-A3B-TIES (589 tokens)
Option A includes bread, jelly, pudding, strawberries, chicken, carrots, garlic, apples, cream. *'Bread' is not correct as 'ladyfingers' were mentioned; 'jelly' is not correct as 'jam' was used; 'strawberries' is incorrect as 'raspberries' were stated; 'chicken' is wrong because 'beef' was used; and other items like 'carrots' and 'garlic' were not mentioned, so A is incorrect.* . . .

Answer: D ✓

Qwen3-Omni-30B-A3B-MI-0.8 (1,728 tokens)
Option A: Bread, jelly, pudding, strawberries, chicken, carrots, garlic, apples, cream - *Bread: she said lady fingers, which are a type of cookie or biscuit, not bread. But lady fingers are often used in trifles, so maybe similar, but not exactly bread. - Jelly: she said jam, which is similar, but jelly is usually fruit spread, jam is thicker. But she said jam. - Pudding: she said custard, which is different. Pudding is usually set with gelatin or cornstarch, custard is egg-based. Not the same. - Strawberries: she said raspberries, not strawberries.* . . .

Answer: D ✓

Qwen3-Omni-30B-A3B-S^3 (238 tokens)
Next, I compared these ingredients with each option: Option A includes bread, strawberries, chicken, carrots, and garlic which were not mentioned; Option B includes cake, marmalade, cherries, turkey, corn, peppers, and melons which were not mentioned; Option C includes cookies, blueberries, pork, beans, tomatoes, peaches, and yogurt which were not mentioned; Option D lists ladyfingers, jam, custard, raspberries, beef, peas, onions, bananas, and whipped cream which exactly matches Speaker 1's description. Therefore, the correct answer is D. . . .

Answer: D ✓

Table 28: **MMAR, Example 19: Counting the Chinese Tones Demonstrated** (70.9% token reduction). S^3 transcribes the last syllable as "má"; the baselines hear "mǎ" and lose the fourth tone.

Problem: How many different Chinese tones are involved in the six Chinese pronunciation words demonstrated by the speaker?
Audio transcript *(reader aid; the model received only the audio)*: "Chinese is easy when you learn four tones in this way. Move your head with the tone shape. Try this."
Ground Truth: 4

Qwen3-Omni-30B-A3B-Thinking (4,147 tokens)
Each syllable's tone is as follows: - First "mā" in "mama": first tone (ā) - Second "mā" in "mama": first tone (ā) - "qí": second tone (á) - "mǎ" in "qí mǎ": third tone (ǎ) *- "mǎ" in "mǎ mā": third tone (ǎ) - "mā" in "mǎ mā": first tone (ā)* The tones used are first, second, and third. *The fourth tone (à) is not demonstrated in any of the six syllables.* Therefore, there are t . . .

Answer: A ×

Qwen3-Omni-30B-A3B-TIES (1,841 tokens)
3. mǎ mā - "mǎ" is third tone. *- "mā" is first tone.* Summary of Tones Used: - First tone: appears in "mā" (repeatedly). - Second tone: appears in "qí". - Third tone: appears in "mǎ" (repeatedly). *The fourth tone is not present in any of the six syllables.*

Answer: A ×

Qwen3-Omni-30B-A3B-MI-0.8 (5,912 tokens)
She says: "ā, á, ǎ, à" – that's the four tones. Then she says "Try this: mā, mā, qí, mǎ, mǎ, mǎ" – let's parse that. She says: "mā, mā, qí, mǎ, mǎ, mǎ" – but that might be a typo or mispronunciation. Looking back: "mā, mā, qí, mǎ, mǎ, mǎ" – but "mā" is repeated, and "qí" is different. Let me read the audio transcript: "Try this: mā, mā, qí, mǎ, mǎ, mǎ" – but it says "mā mā qí mǎ mǎ mǎ" – that seems odd. Perhaps it's "mā, mā, qí, mǎ, mǎ, mǎ" meaning six words. But "mā" is first tone, "qí" is second tone, "mǎ" is third tone, and "mà" might be fourth tone, but she says "mǎ" multiple times. Let's look at the audio: she says "mā, mā, qí, mǎ, mǎ, mǎ" – but "mǎ" is third tone, and she says it three times? That doesn't make sense for different tones.

Answer: A ×

Qwen3-Omni-30B-A3B-S^3 (535 tokens)
Then in the examples: **- mā mā: both tone 1 - qí mǎ: qí is tone 2, mǎ is tone 3 - mǎ má: mǎ is tone 3, má is tone 4 So the tones used are tone 1, tone 2, tone 3, tone 4. All four tones are demonstrated.**

Answer: D ✓

Table 29: **MMSU, Example 20: Types of Disfluency in Speech** (41.8% token reduction). S^3 hears the opening sound as a chuckle; the baselines count it as a filled pause.

Problem: Which types of disfluencies are present in the audio? - Filled pauses: e.g., uh, um - Discourse markers: e.g., well, you know - Restarts: interrupted or repeated sentence starts - Explicit editing terms: e.g., I mean - None: if the speech is fluent.
Audio transcript *(reader aid; the model received only the audio)*: "Well, my sister is the seamstress of the two of us, but she doesn't China paint, so we both have our expertise side."
Ground Truth: discourse markers, restarts

Qwen3-Omni-30B-A3B-Thinking (466 tokens) *Then, the audio starts with a short, sharp laugh: "Heh." This is a common filled pause, often used to indicate a brief hesitation or as a conversational filler. Next, the speaker says: "Well my sister is the seamstress of the two of us..." The word "Well" is a classic discourse marker used to introduce a statement.*	**Qwen3-Omni-30B-A3B-TIES** (1,459 tokens) *First, listen to the speech: "Huh, well my sister is the sinstrous of the two of us. And but she doesn't chain up paint, so we both have our expertise site." Now, break down each type: 1. Filled pauses: The speaker says "Huh" at the beginning. "Huh" is a filled pause, similar to "um" or "uh", used to fill a moment of hesitation.*
Answer: D ×	**Answer:** D ×
Qwen3-Omni-30B-A3B-MI-0.8 (1,944 tokens) *Then, analyzing the audio content: the speaker says, 'Huh. Well, my sister is the seamstress of the two of us, and but she doesn't China paint, so we both have our expertise side.' Specifically, at the very beginning, there's a 'Huh' which is a filled pause (a short, non-lexical sound expressing hesitation or thought), and around 00:03, the speaker says 'and but' which is a classic example of a restart* . . .	**Qwen3-Omni-30B-A3B-S^3** (271 tokens) **First, listen to the beginning: "Ha, well my sister is the simstrous of the two of us..." - The initial "Ha" is a chuckle, which is a filler but not a standard filled pause like "uh" or "um". - The word "well" is a discourse marker, used to introduce the sentence. - The phrase "simstrous" is a self-correction for "simstress". This is a restart/repetition of a word.**
Answer: D ×	**Answer:** A ✓