TO AGGREGATE OR NOT TO AGGREGATE? TEST-TIME AGGREGATION BEYOND VERIFIER-FRIENDLY BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregation-based test-time scaling has produced strong gains on competition mathematics and code generation, but it remains unclear whether those gains transfer beyond verifier-friendly benchmarks, under which task conditions aggregation helps, and when one-step aggregation is sufficient relative to recursive aggregation. We study these questions across structured reasoning, knowledge-intensive reasoning, and medical reasoning, spanning proof-style mathematics, expert-level STEM reasoning, social-science knowledge tasks, BrowseComp-style information seeking, and both tool-free and tool-integrated regimes. Aggregation is effective when sampled trajectories contain recoverably complementary information: complementary reasoning progress in structured reasoning, or complementary retrieved evidence in tool-integrated knowledge/evidence-seeking. Structured reasoning and tool-integrated knowledge/evidence-seeking recover 48% and 56% of available headroom on average, whereas medical reasoning without tools recovers only 21% despite a comparable diversity window. Tool use improves medical base performance, but aggregation remains weak because trajectories more often reflect competing clinical interpretations than composable intermediate progress. Within favorable regimes, aggregation type also matters: RSA is most useful for open-ended proof generation, whereas SSA captures most of the gain in tool-integrated knowledge/evidence-seeking at lower cost. Aggregation value therefore depends jointly on task structure, tool access, and aggregation type rather than following a uniform test-time scaling law across domains.

1 INTRODUCTION

The current wave of agentic AI is increasingly defined by systems that operate through tools and interfaces, including coding agents, deep-research systems, and computer-use agents (Li et al., 2025a; Hong et al., 2023; Singh et al., 2025c; Yao et al., 2023; Zhou et al., 2023). Across these settings, a recurring systems pattern is to sample multiple candidate trajectories, tool-use traces, or search branches and then select or synthesize a final answer. Aggregation-based test-time scaling formalizes this pattern by allocating additional inference-time compute to multiple trajectories and then selecting or combining them (Snell et al., 2024; Wang et al., 2024; Li et al., 2025b; Venkatraman et al., 2025). This approach has produced strong gains on competition mathematics and code generation, where correctness is comparatively easy to verify and partially correct trajectories can often be identified or recombined (Zhang et al., 2025; Singh et al., 2026; Venkatraman et al., 2025). The central open question is whether the same mechanism transfers beyond these verifier-friendly settings.

This paper studies three related questions. First, does aggregation improve performance outside competition math and coding? Second, if it does, under which task conditions does it help? Third, when aggregation is useful, when is a single aggregation pass sufficient and when is recursive aggregation warranted? These questions are domain-dependent. Structured reasoning tasks may admit complementary partial reasoning across trajectories. Tool-integrated knowledge-intensive tasks may benefit from complementary retrieved evidence. medical reasoning often presents a qualitatively different regime in which trajectories encode competing global interpretations rather than composable intermediate progress.

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.

We answer these questions through a cross-domain study of aggregation-based test-time scaling over proof-style mathematics, expert-level STEM reasoning, social-science knowledge tasks, BrowseComp-style information seeking, and medical reasoning. We evaluate open-weight and frontier models in tool-free settings, and open-weight models in tool-integrated settings (Luong et al., 2025; Wei et al., 2025; Center for AI Safety and Scale AI and collaborators, 2025; Zuo et al., 2025; Wu et al., 2025). Our analysis uses two diagnostics: the diversity window, which measures the headroom between Pass@1 and Pass@8, and Aggregation Yield, which measures how much of that headroom an aggregation method recovers.

Our results are domain-specific rather than uniform. Structured reasoning and tool-integrated knowledge/evidence-seeking are the most favorable regimes because independently sampled trajectories often contribute complementary reasoning progress or complementary retrieved evidence. Medical reasoning is the clearest unfavorable regime: tool use substantially improves underlying performance, but aggregation gains remain weak because trajectory diversity is less complementary. Within favorable regimes, aggregation type also matters: SSA is the preferred default in tool-integrated knowledge/evidence-seeking, where a single evidence-aware pass captures most of the gain, whereas RSA is mainly warranted in open-ended proof generation.

Our contributions are:

1. We provide a broad empirical study of aggregation across structured reasoning, knowledge/evidence-seeking, and medical reasoning in tool-free and tool-integrated regimes.
2. We show that aggregation is effective when trajectory diversity is recoverably complementary: complementary reasoning progress in structured reasoning and complementary retrieved evidence in tool-integrated knowledge/evidence-seeking.
3. We identify medical reasoning as the clearest counterexample: substantial multi-sample headroom often fails to convert into aggregation gains because trajectories are less compositionally complementary, even with improved tool-enabled base performance.
4. We distinguish when SSA and RSA are warranted: SSA is the practical default in tool-integrated knowledge/evidence-seeking, whereas RSA is most useful in proof-style structured reasoning.

2 Related Work

Test-time scaling and parallel reasoning. Inference-time scaling spans both sequential depth along a single trajectory and parallel breadth across many trajectories. Recent work shows that search, reranking, and self-verification can substantially improve closed-ended math and coding performance (Snell et al., 2024; Wang et al., 2024; Pan et al., 2025; Lian et al., 2025). Related program-aided methods push on the sequential axis, for example by decomposing Olympiad-style math into executable subproblems (Singh et al., 2025b). Our focus is complementary: rather than proposing a new aggregation method, we ask when parallel aggregation is worthwhile beyond the benchmark regimes on which it is usually evaluated.

Verification and candidate selection. Aggregation depends on deciding which sampled trajectory to trust. Recent work explores verifier-guided search, self-verification, and pairwise comparison as mechanisms for improving candidate selection (Snell et al., 2024; Zhang et al., 2025; Singh et al., 2026). In code generation, stress-testing based filters provide another example of using executable feedback to narrow the gap between Pass@k and final selection (Singh et al., 2025a). These methods are especially effective in verifier-friendly domains. Our study asks how well such selection-based aggregation transfers once trajectories differ not only in correctness but also in evidence, framing, or diagnostic interpretation.

Aggregation and synthesis. Beyond selecting a single candidate, recent work studies whether multiple trajectories can be synthesized into a better final answer. Generative self-aggregation and recursive self-aggregation are motivated by the possibility that different trajectories contain complementary partial reasoning (Li et al., 2025b; Venkatraman et al., 2025). Our study complements this literature by showing that such complementarity is highly domain-dependent.

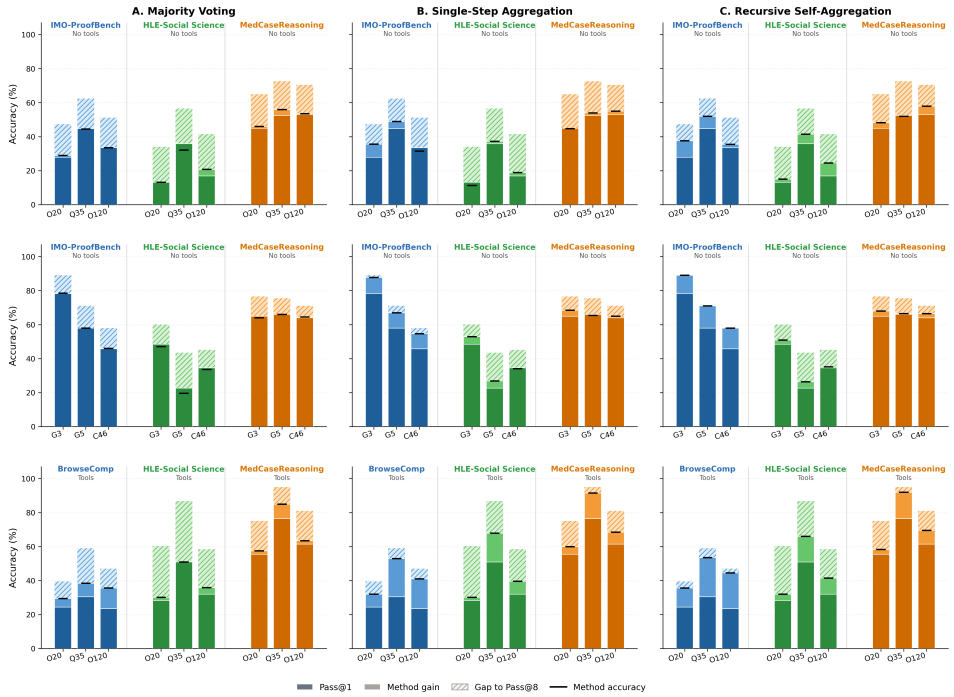


Figure 1: Representative benchmark-level comparison across open-weight no-tool (top), frontier no-tool (middle), and open-weight tool-enabled (bottom) settings. Model proxies are O20 = GPT-OSS-20B, Q35 = Qwen-3.5-35B, O120 = GPT-OSS-120B, G3 = Gemini-3-Pro, G5 = GPT-5, and C46 = Claude 4.6 Opus. ProofBench is the clearest RSA-favorable case, tool-integrated knowledge/evidence-seeking favors SSA, and medical reasoning improves with tools but remains weakly recoverable under aggregation.

Tool-augmented and agentic reasoning. In tool-enabled settings, rollouts differ not only in reasoning quality but also in the evidence they retrieve. Recent work spans ReAct-style reasoning-action loops (Yao et al., 2023; Zhou et al., 2023), deep-research systems (Li et al., 2025a; Singh et al., 2025d), and GUI/computer-use agents (Hong et al., 2023; Singh et al., 2025c; Wei et al., 2025). These settings make aggregation qualitatively different because trajectories can differ in retrieved evidence, action traces, and interface state rather than only in internal reasoning. We study this regime directly by comparing tool-free and tool-integrated aggregation across reasoning, knowledge-seeking, and medical tasks.

3 Experimental Setup

We study aggregation along three axes: *domain*, *model family*, and *tool regime*. This design lets us separate whether aggregation helps at all from which aggregation strategy is warranted in each setting.

3.1 Models and Task Groups

We evaluate open-weight Qwen-3-8B, Qwen-3-30B, Qwen-3.5-35B, GPT-OSS-20B, and GPT-OSS-120B, together with frontier GPT-5, Claude 4.6 Opus, and Gemini-3-Pro. Tool-free experiments use the full set. Tool-integrated experiments focus on GPT-OSS-20B, Qwen-3.5-35B, and GPT-OSS-120B. All frontier-model runs use medium reasoning effort. For compact figure labels, we refer to GPT-OSS-20B, Qwen-3.5-35B, GPT-OSS-120B, Gemini-3-Pro, GPT-5, and Claude 4.6 Opus as O20, Q35, O120, G3, G5, and C46, respectively.

We group tasks into three regimes. **Structured reasoning** includes IMO-ProofBench and HLE-STEM, plus tool-augmented HLE-STEM. **Medical reasoning** includes MedXpertQA and MedCaseReason-

ing. **Knowledge and information-seeking** includes HLE-Social Science without tools and, with tools, HLE-Social Science together with BrowseComp (Luong et al., 2025; Center for AI Safety and Scale AI and collaborators, 2025; Zuo et al., 2025; Wu et al., 2025; Wei et al., 2025). In the analysis, tool-augmented HLE-Social Science and BrowseComp are treated jointly as a knowledge/evidence-seeking regime, whereas tool-augmented HLE-STEM remains grouped with structured reasoning. Dataset subset construction and selected evaluation sizes are detailed in Appendix A.

3.2 Problem Setup

Given an input x and a base model p_θ, we sample $k = 8$ independent worker trajectories

$$\mathcal{T}(x) = \{T_1, \ldots, T_k\}. \tag{1}$$

Each trajectory produces a final answer y_i. In the most general tool-integrated setting,

$$T_i = (x, r_{i,1}, a_{i,1}, o_{i,1}, \ldots, r_{i,m_i}, a_{i,m_i}, o_{i,m_i}, y_i), \tag{2}$$

where $r_{i,j}$ denotes intermediate reasoning, $a_{i,j}$ a tool or environment action, and $o_{i,j}$ the resulting observation. In tool-free settings, the action-observation terms are absent.

The aggregation problem is to map the set of worker trajectories to a single prediction

$$f_\phi(x, \mathcal{T}(x)) \mapsto \hat{y}. \tag{3}$$

We evaluate tool-free and tool-integrated aggregation separately, since in the latter case trajectories differ not only in reasoning but also in the external evidence they uncover.

3.3 Aggregation Methods

We compare four aggregation methods over the sampled worker set $\mathcal{T}(x) = \{T_1, \ldots, T_k\}$.

Majority Voting (MV). As a simple baseline, we select the most frequent final answer among the sampled workers:

$$\hat{y}_{\mathrm{MV}} = \arg\max_y \sum_{i=1}^{k} \mathbf{1}[y_i = y]. \tag{4}$$

MV uses only the final answers and ignores the reasoning traces.

Single-Step Aggregation (SSA). SSA is a single LLM call over all sampled worker trajectories:

$$\hat{y}_{\mathrm{SSA}} = A_\phi(x, \mathcal{T}(x)). \tag{5}$$

SSA is our main one-step aggregation baseline.

Recursive Self-Aggregation (RSA). RSA repeatedly aggregates randomly sampled subsets. Let K be the subset size, N the number sampled per round, and P the number of rounds. Starting from $\mathcal{P}^{(0)} = \mathcal{T}(x)$, round $p = 0, \ldots, P-1$ samples

$$S_1^{(p)}, \ldots, S_N^{(p)} \subseteq \mathcal{P}^{(p)}, \qquad |S_j^{(p)}| = K. \tag{6}$$

$$\mathcal{P}^{(p+1)} = \left\{ A_\phi\left(x, S_j^{(p)}\right) \right\}_{j=1}^{N}. \tag{7}$$

After P rounds, one final SSA call produces

$$\hat{y}_{\mathrm{RSA}} = A_\phi\left(x, \mathcal{P}^{(P)}\right). \tag{8}$$

This repeated reduction is intended to synthesize complementary candidate trajectories (Venkatraman et al., 2025).

Compressed Evidence Aggregation (CEA). In tool-integrated settings, full trajectories are often too long to aggregate directly. Our main tool-based method is **Compressed Evidence Aggregation (CEA)**: for each worker T_i with final answer y_i, we select only the answer-relevant tool evidence and the final reasoning step, constructing

$$e_i = E_\phi(x, T_i, y_i), \tag{9}$$

where e_i contains the selected evidence and r_i^{final} denotes the final reasoning step immediately preceding y_i. Aggregation is then performed over the compressed worker representations:

$$\hat{y}_{\text{CEA}} = A_\phi\big(x, \{(e_i, r_i^{\text{final}}, y_i)\}_{i=1}^k\big)\,. \tag{10}$$

We use CEA in all main tool-integrated experiments.

3.4 Evaluation Protocol

For each model-task pair, we sample $k = 8$ independent worker trajectories per problem using temperature-based decoding. For the open-weight reasoning models, we use the family-recommended temperatures: 0.7 for GPT-OSS models and 1.5 for Qwen reasoning models. We report Pass@1, Pass@8, and post-aggregation accuracy for each method. On the classification benchmarks, Pass@8 measures whether at least one of the k workers is correct and therefore serves as an oracle upper bound on recoverable performance from the sampled worker pool. IMO-ProofBench instead uses graded LLM-judge scores, so its Pass@8 and MV entries are proxy values: Pass@8 is the maximum score assigned by the judge among the eight worker proofs, and MV is the mode of the eight judge-assigned scores. All other benchmarks are judged with an LLM exact-match-style scorer. IMO-ProofBench uses rubric-based proof evaluation. Judge details are provided in Appendix B. We also track the total token cost of each method, including worker generation and all aggregation calls. For CEA this includes both evidence compression and the final aggregation call.

All reported scores are averages across three runs. The $\pm$ value alongside each cell in Tables 2 to 5 is one paired problem-level standard error over the per-problem outcomes.

4 Analysis

We analyze aggregation performance using two quantities: the diversity window, defined as Pass@8 $-$ Pass@1, and Aggregation Yield (AY), which measures the fraction of this headroom recovered by a given aggregation method. The central distinction is between raw multi-sample headroom and whether the resulting trajectory variation is recoverably complementary. These quantities are used throughout the section to compare aggregation behavior across domains, model families, and tool regimes.

4.1 Diversity Window and Aggregation Yield

For each model-task pair, we sample $k = 8$ worker trajectories and report Pass@1, Pass@8, and post-aggregation accuracy. Pass@1 measures the expected accuracy of a single sampled worker. On classification benchmarks, Pass@8 measures whether at least one of the k sampled workers is correct. The IMO-ProofBench proxy is defined in the evaluation protocol above. The difference

$$\text{Pass@8} - \text{Pass@1}$$

defines the **Diversity window (DW)**. It measures the amount of recoverable headroom available in the sampled worker pool.

Raw aggregation gains are not directly comparable across tasks with different amounts of available headroom. We therefore define the **Aggregation Yield (AY)** of a method $\mathcal{S}$ as

$$\text{AY}(\mathcal{S}) = \frac{\text{Acc}(\mathcal{S}) - \text{Pass@1}}{\text{Pass@8} - \text{Pass@1}},$$

where $\text{Acc}(\mathcal{S})$ is the final accuracy after applying aggregation method $\mathcal{S}$. AY is equal to 0 when aggregation provides no gain over a single sample, equal to 1 when it fully recovers the available headroom, and negative when aggregation degrades performance. For readability we report AY as

Table 1: Regime summary. DW = Pass@8 − Pass@1. AY = recovered headroom.

	No tools		Tools	
Regime	**DW**	**AY (%)**	**DW**	**AY (%)**
Structured	16.44	48	33.33	42
Knowledge	19.49	21	26.91	56
Medical	15.77	21	22.25	37

the percentage of recoverable headroom captured by the best of {MV, SSA, RSA} per row (which positively biases the aggregate relative to any single method. Per-method AY can be recomputed from Tables 2 to 4). AY distinguishes cases in which aggregation fails because little headroom exists from cases in which substantial headroom exists but is not exploitable.

4.2 Aggregation Depends on Complementarity, Not Headroom Alone

The main empirical distinction is not the size of the diversity window by itself, but whether the induced trajectory diversity is complementary and recoverable. Table 1 summarizes the regime-level pattern. Structured reasoning without tools has average DW of 16.44 pp and average AY of 48% (61% on IMO-ProofBench vs. 36% on HLE-STEM, indicating the average is driven predominantly by proof-style tasks). Tool-augmented HLE-STEM, treated as a structured variant, has average DW of 33.33 pp and average AY of 42%. The strongest recoverability appears in tool-integrated knowledge/evidence-seeking, which has average DW of 26.91 pp and average AY of 56%. By contrast, knowledge-intensive reasoning without tools and medical reasoning without tools each recover only 21% of available headroom despite non-trivial diversity windows. Medical reasoning with tools improves to 37%, but remains well below the knowledge/evidence-seeking tool regime.

These contrasts are consistent with distinct forms of trajectory variation. In structured reasoning tasks, sampled trajectories often contain different partial advances toward the same solution. In no-tool knowledge-intensive settings, performance is often bottlenecked by missing or weakly represented facts, limiting what aggregation can recover. In tool-integrated knowledge-intensive settings, trajectories differ in the evidence they retrieve, so aggregation benefits from selecting and consolidating the most informative evidence-bearing rollouts. In medical reasoning, by contrast, trajectories more often represent competing global interpretations, making disagreement less amenable to aggregation.

The main methodological implication is that raw headroom is not sufficient to predict aggregation gains. Benchmarks with similar Pass@8 minus Pass@1 gaps can have substantially different post-aggregation outcomes because the underlying trajectory diversity is structurally different.

4.3 Structured Reasoning: Complementary Reasoning Progress Favors RSA

Structured reasoning is the most favorable no-tool regime because sampled trajectories often contain compatible partial derivations, lemmas, or decompositions that can be identified or recombined. Across the structured no-tool rows in Table 1, average AY is 48%. On HLE-STEM, aggregation yields large improvements for frontier models: GPT-5 improves from 38.67% to 50.22%, and Claude 4.6 Opus improves from 42.67% to 58.22%.

IMO-ProofBench is more favorable still to recursive aggregation. RSA has the highest point estimate in every ProofBench row in Table 2. RSA improves GPT-OSS-20B from 27.78% to 37.78%, GPT-5 from 57.78% to 71.11%, and Claude 4.6 Opus from 45.56% to 57.78%. The contrast with HLE-STEM is informative: SSA captures most of the available gain there and exceeds RSA for Gemini-3-Pro and GPT-5, whereas RSA has consistently higher point estimates on IMO-ProofBench. This pattern is consistent with proof generation placing greater value on combining reasoning progress distributed across multiple trajectories.

Tool access does not alter the basic character of HLE-STEM. Across the three tool-augmented HLE-STEM rows in Table 5, average DW is 33.33 pp and average AY is 42%. Both SSA and RSA improve substantially over Pass@1, and the RSA vs SSA gap is within one standard error and is not statistically resolved on this benchmark. Consistent with this, on no-tool HLE-STEM RSA is below SSA for Gemini-3-Pro and GPT-5 (Table 2). The RSA advantage in structured reasoning is therefore

Table 2: **No-tool results on HLE-STEM, HLE-Social Science, and IMO-ProofBench.** Best aggregation result per row is shown in **bold**. Values are scores (%). For IMO-ProofBench, Pass@8 and MV are proxy quantities defined via the graded proof judge (max and mode of the eight rubric scores, see §4.1).

	HLE-STEM					HLE-Social Science					IMO-ProofBench				
Model	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA
Qwen-3-8B	1.33 ±1.35	12.00 ±3.81	1.33 ±1.34	**3.11** ±2.02	**3.11** ±2.06	3.80 ±2.61	30.22 ±6.42	11.40 ±4.39	**13.13** ±4.73	**13.13** ±4.69	19.44 ±3.14	33.89 ±3.96	20.56 ±3.42	23.89 ±3.58	**25.00** ±3.74
GPT-OSS-20B	12.00 ±3.77	32.44 ±5.46	9.33 ±3.42	13.33 ±3.92	**19.11** ±4.55	13.13 ±4.63	34.02 ±6.56	13.13 ±4.65	11.40 ±4.38	**15.03** ±4.84	27.78 ±4.50	47.22 ±4.85	28.89 ±4.43	35.56 ±4.84	**37.78** ±4.83
Qwen-3-30B	8.00 ±3.16	23.11 ±4.87	6.67 ±2.90	8.00 ±3.21	**12.44** ±3.85	11.40 ±4.35	32.12 ±6.44	9.50 ±3.98	9.50 ±4.02	**11.40** ±4.35	27.78 ±3.94	38.33 ±4.44	25.56 ±4.02	29.44 ±4.50	**32.22** ±4.46
Qwen-3.5-35B	36.44 ±5.60	55.56 ±5.79	36.44 ±5.63	35.56 ±5.75	**36.89** ±5.70	35.92 ±6.52	56.65 ±6.73	32.12 ±6.31	37.31 ±6.69	**41.45** ±6.73	45.00 ±4.12	62.22 ±4.56	44.44 ±4.20	48.89 ±4.42	**52.22** ±4.48
GPT-OSS-120B	20.00 ±4.66	40.44 ±5.73	15.11 ±4.19	21.78 ±4.75	**24.44** ±5.02	16.93 ±5.21	41.45 ±6.87	20.73 ±5.59	18.83 ±5.43	**24.53** ±5.96	33.33 ±4.33	51.11 ±4.68	33.33 ±4.37	31.67 ±4.34	**35.56** ±4.50
Gemini-3-Pro	61.78 ±4.49	82.22 ±4.44	64.89 ±5.60	**71.11** ±5.25	68.00 ±5.41	48.53 ±3.27	60.10 ±3.53	47.15 ±3.59	**53.02** ±3.61	50.95 ±3.61	78.33 ±3.68	88.89 ±2.47	78.33 ±4.07	87.78 ±3.35	**88.89** ±4.01
GPT-5	38.67 ±4.63	60.89 ±5.66	33.78 ±5.51	**50.22** ±5.97	47.11 ±5.94	22.80 ±2.52	43.52 ±3.57	19.69 ±2.88	**26.94** ±3.22	26.42 ±3.21	57.78 ±4.81	71.11 ±4.19	57.78 ±5.59	67.22 ±5.16	**71.11** ±5.22
Claude 4.6 Opus	42.67 ±5.21	61.78 ±5.89	42.67 ±6.04	54.22 ±6.07	**58.22** ±6.03	34.72 ±3.09	45.08 ±3.59	33.68 ±3.38	34.20 ±3.42	**35.23** ±3.45	45.56 ±4.15	57.78 ±4.25	46.11 ±4.85	54.44 ±4.95	**57.78** ±4.94

Table 3: **No-tool medical reasoning.** Best result in **bold**. Values are scores (%).

	MedXpertQA					MedCaseReasoning				
Model	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA
Qwen-3-8B	16.50 ±2.63	33.50 ±3.31	**20.50** ±2.83	18.50 ±2.75	20.00 ±2.82	30.50 ±3.25	49.50 ±3.52	35.00 ±3.38	**35.83** ±3.40	35.67 ±3.41
GPT-OSS-20B	30.50 ±3.27	56.00 ±3.50	34.00 ±3.37	**35.50** ±3.39	33.50 ±3.35	45.00 ±3.53	65.00 ±3.39	46.00 ±3.51	44.67 ±3.55	**48.17** ±3.49
Qwen-3-30B	26.00 ±3.12	40.00 ±3.48	**27.50** ±3.18	23.50 ±3.00	24.50 ±3.04	36.50 ±3.40	56.00 ±3.56	38.50 ±3.46	37.50 ±3.41	**40.00** ±3.46
Qwen-3.5-35B	40.50 ±3.48	60.00 ±3.47	**47.50** ±3.56	46.00 ±3.55	41.17 ±3.52	52.50 ±3.56	72.50 ±3.19	**56.00** ±3.53	54.00 ±3.57	52.00 ±3.59
GPT-OSS-120B	28.50 ±3.21	57.50 ±3.48	32.50 ±3.31	38.50 ±3.48	**39.50** ±3.47	53.00 ±3.53	70.50 ±3.22	53.50 ±3.51	55.00 ±3.49	**58.00** ±3.46
Gemini-3-Pro	71.00 ±2.96	78.00 ±2.88	70.00 ±3.23	71.00 ±3.19	**73.00** ±3.12	65.00 ±3.05	76.50 ±3.04	64.00 ±3.41	**68.50** ±3.31	68.00 ±3.31
GPT-5	57.00 ±3.24	66.00 ±3.38	55.00 ±3.53	56.00 ±3.51	**56.50** ±3.54	66.00 ±3.14	75.50 ±3.07	66.00 ±3.40	65.50 ±3.40	**66.50** ±3.39
Claude 4.6 Opus	51.83 ±3.34	59.00 ±3.47	51.00 ±3.51	**52.00** ±3.51	50.50 ±3.51	64.00 ±3.23	71.00 ±3.19	64.50 ±3.36	65.00 ±3.33	**66.50** ±3.34

driven predominantly by IMO-ProofBench rather than by HLE-STEM. Even in tool-augmented structured reasoning, most recoverable gain is captured by a single aggregation pass. Recursion is most clearly justified in proof-style structured reasoning.

Tools improve medical reasoning, but they do not change its qualitative position. Across the six open-weight medical settings with tools, average AY rises from 21% to 37%, still well below the 56% observed for the knowledge-intensive tool regime. Table 4 shows large gains on MedCaseReasoning, especially for Qwen-3.5-35B and GPT-OSS-120B, but medical trajectories remain markedly less aggregation-friendly than tool-integrated knowledge/evidence-seeking.

4.4 To aggregate or not to aggregate?

Different favorable regimes do not call for the same aggregation strategy. Signed RSA-minus-SSA deltas are largest and most consistent in structured reasoning without tools, driven primarily by ProofBench. This is the clearest setting in which recursive aggregation recovers complementary reasoning progress that a single aggregation pass leaves unused. Figure 2 summarizes the contrast across regimes.

4.5 Knowledge and Evidence-Seeking: Complementary Evidence Favors SSA

Knowledge-intensive reasoning splits cleanly into two regimes. Without tools, HLE-Social Science is a boundary condition for aggregation: average DW is 19.49 pp, but average AY is only 21% (Table 1). Gains exist, but they remain modest relative to available headroom. For example, GPT-5 improves from 22.80% to 26.94%, and Claude 4.6 Opus improves only from 34.72% to 35.23%. The limiting

Table 4: **Tool-integrated medical reasoning.** CEA aggregation. Best result in **bold**. Values are scores (%).

Model	MedXpertQA					MedCaseReasoning				
	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA
GPT-OSS-20B	38.50 ±3.46	67.50 ±3.32	**43.00** ±3.49	39.50 ±3.45	41.00 ±3.48	55.50 ±3.45	75.00 ±3.04	57.50 ±3.50	**60.00** ±3.45	58.33 ±3.49
Qwen-3.5-35B	41.50 ±3.50	66.00 ±3.37	**49.00** ±3.54	48.00 ±3.54	48.00 ±3.55	76.50 ±3.00	95.00 ±1.54	85.00 ±2.81	91.50 ±1.97	**92.00** ±1.92
GPT-OSS-120B	44.00 ±3.50	66.50 ±3.33	47.00 ±3.56	49.00 ±3.55	**50.00** ±3.57	61.50 ±3.42	81.00 ±2.77	63.50 ±3.38	68.50 ±3.28	**69.50** ±3.24

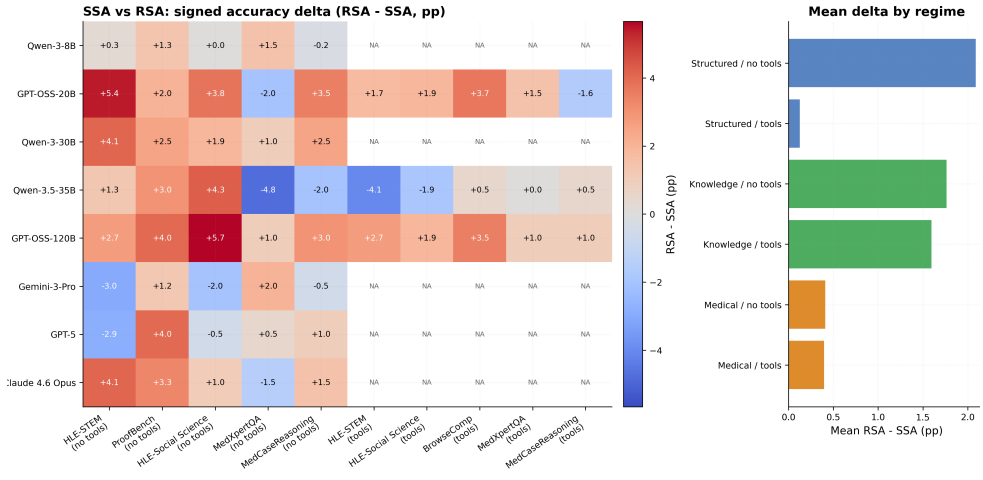


Figure 2: **Signed RSA-minus-SSA deltas across regimes.** Positive values favor RSA. Recursive aggregation is most consistently helpful in structured reasoning without tools, while tool-integrated knowledge/evidence-seeking and medical reasoning remain close to zero, favoring SSA as the default operating point.

factor is factual coverage: when the relevant fact is absent or only weakly represented, repeated sampling does not create the kind of complementary structure that aggregation can reliably exploit.

The picture changes once evidence acquisition is enabled. Treating tool-augmented HLE-Social Science together with BrowseComp as a single knowledge/evidence-seeking tool regime yields average DW of 26.91 pp and average AY of 56%, the highest value in Table 1. BrowseComp provides the clearest illustration: the best aggregation result improves over Pass@1 by 11.17 pp for GPT-OSS-20B, 23.00 pp for Qwen-3.5-35B, and 21.00 pp for GPT-OSS-120B. Tool-augmented HLE-Social Science shows the same qualitative shift, especially for Qwen-3.5-35B, which improves from 50.95% to 67.88%. The underlying mechanism is complementary retrieved evidence: different trajectories surface different relevant snippets, and aggregation succeeds by consolidating them.

Across these six knowledge-intensive tool rows, both SSA and RSA improve substantially over Pass@1 (by 12.29 pp and 13.89 pp on average respectively). The RSA vs SSA gap of 1.59 pp is within one standard error and should be read as null on the current test sets. Most of the available gain therefore arises from a single evidence-aware aggregation step rather than from deeper recursive synthesis. Accordingly, SSA is the preferred operating point in this regime: RSA can add small increments, but the principal bottleneck is evidence discovery and retention rather than the recombination of partially correct internal reasoning.

4.6 Medical Reasoning: Tools Help, Aggregation Remains Weak

Medical reasoning is the clearest counterexample. These benchmarks often exhibit substantial headroom, but much less of it is recoverable by aggregation. Across the no-tool medical rows, average DW is 15.77 pp, close to the structured-reasoning average, but average AY is only 21% (Table 1).

Table 5: **Tool-integrated results on HLE-STEM, HLE-Social Science, and BrowseComp.** Tool-based aggregation uses CEA by default. Best aggregation result per row is shown in **bold**. Values are scores (%).

Model	HLE-STEM					HLE-Social Science					BrowseComp				
	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA	P@1	P@8	MV	SSA	RSA
GPT-OSS-20B	17.78 ±4.41	51.56 ±5.76	25.78 ±5.08	27.11 ±5.16	**28.89** ±5.28	28.32 ±6.09	60.45 ±6.65	30.22 ±6.33	30.22 ±6.34	**32.12** ±6.44	24.50 ±3.07	39.50 ±3.43	29.50 ±3.22	32.00 ±3.31	**35.67** ±3.52
Qwen-3.5-35B	29.78 ±5.23	66.22 ±5.57	39.11 ±5.64	**47.11** ±5.82	43.11 ±5.75	50.95 ±6.87	86.87 ±4.62	50.95 ±6.90	**67.88** ±6.42	65.98 ±6.52	30.50 ±3.26	59.00 ±3.51	38.50 ±3.48	53.00 ±3.53	**53.50** ±3.56
GPT-OSS-120B	24.44 ±5.03	54.22 ±5.85	32.00 ±5.49	35.11 ±5.53	**37.78** ±5.62	32.12 ±6.47	58.55 ±6.85	35.92 ±6.56	39.55 ±6.74	**41.45** ±6.77	23.50 ±3.02	47.00 ±3.50	35.67 ±3.39	41.00 ±3.43	**44.50** ±3.49

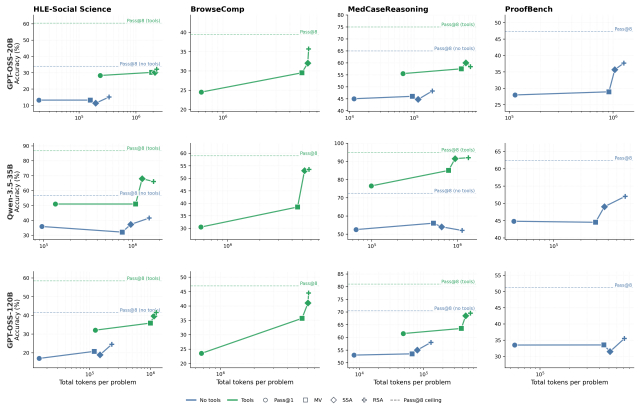


Figure 3: Accuracy-cost Pareto curves for the open-weight models. RSA most clearly extends the frontier on ProofBench. On HLE-Social Science, BrowseComp, and MedCaseReasoning, tool use dominates and SSA captures most practical gain.

On MedXpertQA, Qwen-3.5-35B has a 19.50 pp diversity window (40.50% to 60.00%) but reaches only 47.50% after aggregation, and GPT-5 remains below Pass@1 after aggregation despite Pass@8 of 66.00%. Majority voting is also unusually competitive on MedXpertQA, consistent with medical trajectories differing more often in global diagnostic framing than in separable intermediate steps. Figure 2 identifies structured reasoning as the main regime in which recursion changes the outcome rather than merely adding cost.

4.7 Accuracy-Cost Tradeoffs

Because SSA aggregates all eight workers, its total cost is nearly an order of magnitude above Pass@1 (about $9.8\times$ without tools, $9.3\times$ with tools in our open-weight runs). RSA then adds $1.6\times$ (no tools) or $1.2\times$ (tools) over SSA. The question is therefore where the added cost actually shifts the accuracy-cost frontier (Figure 3).

Three frontier patterns emerge. First, tool use shifts the frontier more than deeper aggregation when success depends on retrieved evidence (HLE-Social Science, BrowseComp). Second, once complementary evidence has been surfaced, SSA is the preferred operating point. RSA adds little across those tasks and tool-enabled MedCaseReasoning. Third, recursion matters most for proof-style reasoning: on ProofBench, RSA raises Qwen-3.5-35B from 48.89% to 52.22% and GPT-OSS-120B from 31.67% to 35.56%.

5 Conclusion

Aggregation transfers beyond verifier-friendly benchmarks only when sampled trajectories are complementary: structured reasoning benefits from compatible reasoning progress, knowledge/evidence-seeking from complementary retrieved evidence, and medical reasoning remains unfavorable even with tools. Thus, use tools first when evidence matters, prefer SSA for knowledge/evidence-seeking, and reserve RSA for proof-style tool-free structured reasoning (IMO-ProofBench in our study).

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, or interpreting results, and the remaining required-disclosure tasks of generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code and editing a research paper to improve readability. We have reviewed all AI-assisted work. AI-generated code was verified and tested for correctness by the authors, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Center for AI Safety and Scale AI and collaborators. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2023. URL https://arxiv.org/abs/2312.08914.

Baixuan Li, Dingchu Zhang, Jialong Wu, Wenbiao Yin, Zhengwei Tao, Yida Zhao, Liwen Zhang, Haiyang Shen, Runnan Fang, Pengjun Xie, Jingren Zhou, and Yong Jiang. Parallelmuse: Agentic parallel thinking for deep information seeking. *arXiv preprint arXiv:2510.24698*, 2025a.

Zichong Li, Xinyu Feng, Yuheng Cai, Zixuan Zhang, Tianyi Liu, Chen Liang, Weizhu Chen, Haoyu Wang, and Tuo Zhao. Llms can generate a better answer by aggregating their own responses. *arXiv preprint arXiv:2503.04104*, 2025b.

Long Lian, Sida Wang, Felix Juefei-Xu, Tsu-Jui Fu, Xiuyu Li, Adam Yala, Trevor Darrell, Alane Suhr, Yuandong Tian, and Xi Victoria Lin. Threadweaver: Adaptive threading for efficient parallel reasoning in language models. *arXiv preprint arXiv:2512.07843*, 2025.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning. *arXiv preprint arXiv:2511.01846*, 2025.

Y. Pan et al. Learning adaptive parallel reasoning with language models. *arXiv preprint arXiv:2504.15466*, 2025.

Harman Singh et al. V1: Unifying generation and self-verification for parallel reasoners. https://harmandotpy.github.io/v1-verification/, 2026. Project page.

Kunal Singh, Sayandeep Bhowmick, Pradeep Moturi, and Siva Kishore Gollapalli. No stress no gain: Stress testing based self-consistency for olympiad programming. Published at the ICLR 2025 VerifAI: AI Verification in the Wild, 2025a. URL https://openreview.net/pdf?id=7SlCSjhBsq.

Kunal Singh, Ankan Biswas, Sayandeep Bhowmick, Pradeep Moturi, and Siva Kishore Gollapalli. Sbsc: Step-by-step coding for improving mathematical olympiad performance, 2025b. URL https://arxiv.org/abs/2502.16666.

Kunal Singh, Shreyas Singh, and Mukund Khanna. Trishul: Towards region identification and screen hierarchy understanding for large vlm based gui agents, 2025c. URL https://arxiv.org/abs/2502.08226.

Shreyas Singh, Kunal Singh, and Pradeep Moturi. Fathom-deepresearch: Unlocking long horizon information retrieval and synthesis for slms, 2025d. URL https://arxiv.org/abs/2509.24107.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

S. Venkatraman et al. Recursive self-aggregation unlocks deep thinking in large language models. *arXiv preprint arXiv:2509.26626*, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Fuxiang Zhang, Jiacheng Xu, Chaojie Wang, Ce Cui, Yang Liu, and Bo An. Incentivizing llms to self-verify their answers. *arXiv preprint arXiv:2506.01369*, 2025.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models, 2023. URL https://arxiv.org/abs/2310.04406.

Yizhi Zuo et al. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Dataset Subset Construction

HLE subsets. HLE-STEM and HLE-Social Science are derived from HLE using its LLM-as-a-judge verification annotations. We retain text-only Gold-subset items for which the problem, reference answer, and rationale are all marked valid. HLE-STEM pools the Physics, Chemistry, and Engineering categories, yielding 74 items in our evaluation. The HLE Humanities/Social Science subset yields 53 items in our evaluation.

Other evaluation subsets. IMO-ProofBench contains 60 problems. The medical evaluation uses fixed 200-example subsets of MedXpertQA and the MedCaseReasoning test split, and BrowseComp uses a fixed 200-example subset. These are computationally tractable representative slices rather than exhaustive full-benchmark evaluations.

Effective per-benchmark N used throughout Tables 2 to 4: HLE-STEM $= 74$, HLE-Social Science $= 53$, IMO-ProofBench $= 60$, MedXpertQA $= 200$, MedCaseReasoning $= 200$, BrowseComp $= 200$.

B Experimental Details

Aggregator and evidence extractor. Throughout the paper, both the aggregator A_ϕ (Eqs. 5, 8, 10) and the CEA evidence extractor E_ϕ (Eq. 9) are instantiated with *the same model as the worker being evaluated* in every row. That is, every reported number corresponds to a self-aggregation and self-extraction setting, so differences across rows reflect differences in the base model rather than in the aggregator or extractor identity.

Judge. We use Gemini-3.1-Pro as the LLM judge for all evaluations that require model-based adjudication. It performs exact-match-style answer-equivalence judgments for the non-proof benchmarks and rubric-based proof-quality judgments for IMO-ProofBench.

Worker decoding. Open-weight workers use family-recommended sampling temperatures (0.7 for GPT-OSS models and 1.5 for Qwen reasoning models), as also noted in §3.3. Frontier workers (GPT-5, Claude 4.6 Opus, Gemini-3-Pro) use the "medium" reasoning effort setting exposed by each vendor's API.

RSA configuration. Recursive Self-Aggregation uses subset size $K = 4$, $N = 8$ subsets per round, and $P = 3$ rounds followed by a final aggregation call. The same (K, N, P) is used across all rows.

Tool environment. Tool-integrated experiments expose a web-search interface backed by the Google Search API and the Serper API. The same tool schema is used across all tool-integrated rows.

C Qualitative Examples

We present three representative case studies. Each block includes an abridged question, representative trajectories, the aggregation outcome, and the interpretation that connects the example back to the main argument.

Case A. Structured reasoning: complementary proof progress is recoverable **ProofBench `q000011`, GPT-OSS-20B**

Question (abridged). Choose 1000 integers from $\{1, \ldots, 2000\}$ so that no chosen number divides another chosen number, and determine the smallest possible minimum element.

Raw pool. The eight sampled proofs are weak overall: 7 are judged wrong, 1 is judged partial, and the average normalized proof score is 0.0875. No single rollout contains a complete correct proof.

Representative trajectories.

- **p0001 (partial lower-bound argument).** Argues that if $m_A \leq 666$, then at least two numbers in $(1000, 2000]$ are divisible by m_A, and therefore concludes $m_A \geq 667$.
- **p0004 (incorrect constructive argument).** Introduces a chain-based decomposition of the divisibility poset, but the construction is flawed and the final answer is $m_A = 500$.
- **SSA.** Remains weak in this case: it effectively follows trajectory `p0001`, and the resulting proof is still judged wrong with normalized score 0.1.
- **RSA.** Synthesizes the lower-bound idea with an explicit construction excluding the critical multiple 1334, producing an almost-correct proof with normalized score 0.7.

Interpretation. This is the clearest qualitative case in which the useful information is distributed across trajectories. One proof contains the right lower bound, another contains part of the constructive idea, and recursive aggregation is able to combine them into a much stronger proof. This is precisely the regime in which RSA is warranted.

Case B. Tool-integrated knowledge/evidence-seeking: one evidence-bearing rollout is enough **BrowseComp `q000036`, Qwen-3.5-35B**

Question (abridged). Connect a November 2019 article about scoring methods in one sport to a January 2020 interview article about the same athlete, then identify that athlete's hometown.

Raw pool. Only one of the eight tool-using trajectories is correct, so the candidate accuracy is 0.125. The useful evidence is extremely sparse: the correct answer is supported by a single evidence-bearing path, while several other rollouts cluster around plausible but unsupported athletes.

Representative trajectories.

- **p0001 (wrong athlete cluster).** Interprets the query through a Patrick Mahomes path and answers *Tyler, Texas*.
- **p0002 (evidence-bearing trajectory).** Retrieves the relevant article pair and answers *Mallorca, Spain*. Its tool summary identifies a single decisive evidence item supporting that answer.
- **p0003 / p0006 (another wrong cluster).** Converge on Shaun White and answer *Carlsbad, California*.
- **SSA.** Selects the evidence-bearing trajectory and outputs *Mallorca, Spain*. The key point is that once the relevant evidence is present in the pool, a single aggregation step is sufficient.

Interpretation. This example shows why SSA is the preferred operating point in tool-integrated knowledge/evidence-seeking. Once a single rollout has retrieved the right evidence, the main challenge is simply to surface it. One evidence-bearing trajectory can already be enough.

Case C. Medical reasoning: disagreement is not complementary MedCaseReasoning `q000037`, Qwen-3.5-35B

Question (abridged). A 58-year-old woman presents with chronic necrotizing lymphadenitis, nasal chondritis, conjunctivitis, vasculitic skin lesions, persistent fevers, and positive *Bartonella henselae* serology. Identify the most likely diagnosis.

Raw pool. The candidate pool is not sparse: 4 of the 8 trajectories are correct, so the candidate accuracy is 0.5. The issue is not absence of correct trajectories, but disagreement between two coherent global interpretations of the case.

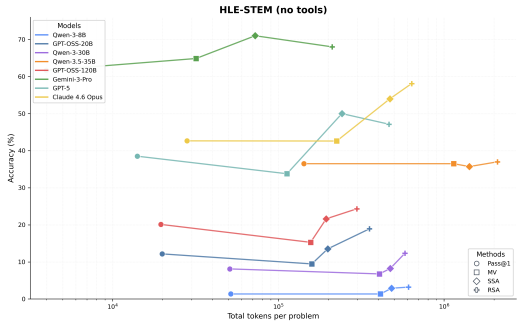
Representative trajectories.

- **p0001 (infectious framing).** Leans on the chronic lymphadenitis and serology and answers *Bartonella henselae infection (cat-scratch disease)*.
- **p0002 (autoimmune framing).** Treats nasal chondritis as pathognomonic and commits to *Relapsing Polychondritis*.
- **p0004 / p0006.** Again identify *Bartonella henselae infection*, showing that correct rollouts are present multiple times in the raw pool.
- **SSA.** Outputs *Relapsing Polychondritis*, following the wrong autoimmune interpretation.
- **RSA.** Also outputs *Relapsing Polychondritis*. Recursion does not rescue the error.

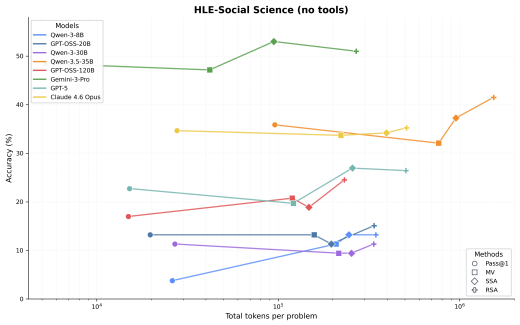
Interpretation. This is the clearest qualitative counterexample in the paper. The raw pool already contains several correct diagnoses, but the disagreement is not decomposable into complementary intermediate pieces. Instead, aggregation amplifies the wrong global interpretation, which is exactly why medical reasoning remains less aggregation-friendly than structured reasoning or tool-integrated evidence-seeking.

D FULL PARETO FRONTIERS

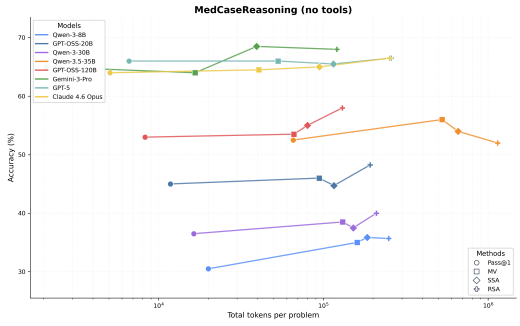
We include the full per-dataset Pareto frontiers used to support the accuracy-cost analysis in the main paper. Each panel plots Pass@1, MV, SSA, and RSA against total tokens per problem. Dashed horizontal lines indicate the Pass@8 ceiling for the corresponding condition.



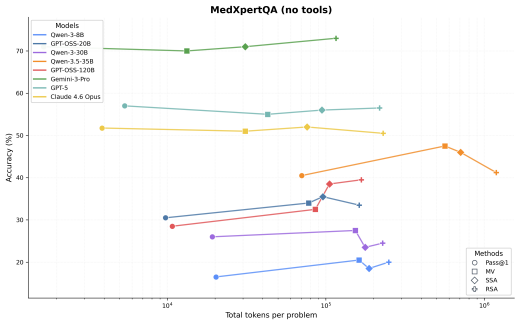
(a) HLE-STEM (no tools)



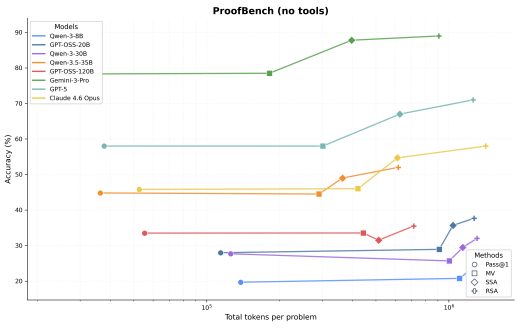
(b) HLE-Social Science (no tools)



(c) MedCaseReasoning (no tools)

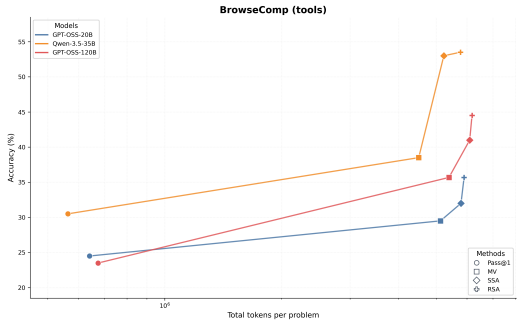


(d) MedXpertQA (no tools)

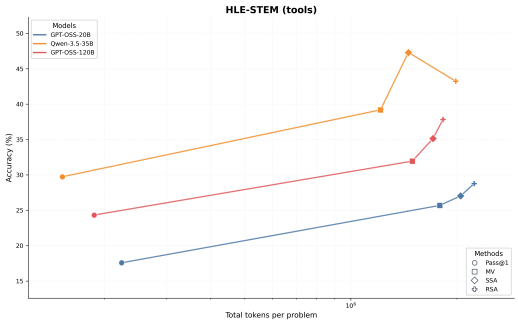


(e) ProofBench (no tools)

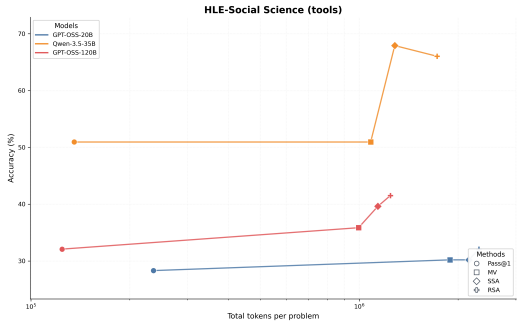
Figure 4: **Full per-dataset Pareto frontiers: no-tool settings.**



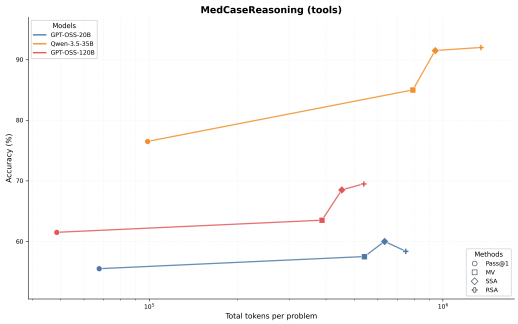
(a) BrowseComp (tools)



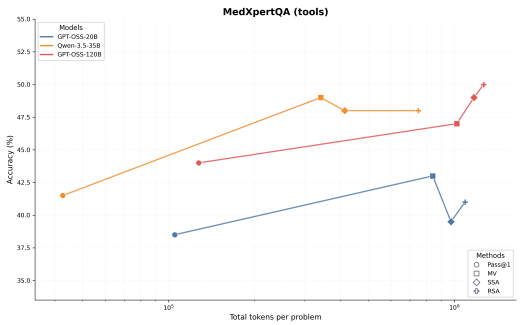
(b) HLE-STEM (tools)



(c) HLE-Social Science (tools)



(d) MedCaseReasoning (tools)



(e) MedXpertQA (tools)

Figure 5: **Full per-dataset Pareto frontiers: tool-integrated settings.**

E AGGREGATOR SYSTEM PROMPTS

We reproduce the exact system prompts used by the aggregation model. We report the generic answer-selection and answer-synthesis prompts used in the main no-tool runs, the tool-integrated variants used over CEA summaries, and the proof-specific variants used on IMO-ProofBench.

E.1 SSA SYSTEM PROMPTS

SSA selector prompt (no tools).

```
You are a final answer selector.

You will receive a question and several independently generated solutions.
Each solution contains a trajectory id, a final answer, and a reasoning trace.

Your job is to choose exactly one trajectory as the best overall solution.
Base the decision on likely correctness, not fluency.

How to judge:
- First evaluate each solution independently on its own merits.
- Ask whether its answer and reasoning actually fit the details of the
   question.
- Reward faithful, coherent, evidence-grounded reasoning.
- Penalize unsupported leaps, contradictions, vague answers, and shallow
   pattern-matching.
- Then use cross-solution comparison as evidence:
   independent agreement across multiple trajectories can increase confidence,
   especially when the reasoning converges for similar underlying reasons.
- Do not treat majority vote as decisive. A minority solution can still be
   best if its reasoning better explains the question.
- Distinguish genuine self-consistency from repeated but weak or circular
   reasoning or unbased assumptions.
- Do not reward verbosity, confidence, or stylistic polish.
- Do not solve the question from scratch.
- Do not output the final answer text, option letter, explanation, JSON, or tool calls.

- Once you are done thinking, Your output must exactly contain one tag of this form:
<trajectory_id>p0006</trajectory_id>
```

SSA selector prompt (tool-integrated / CEA).

```
You are a final answer selector for tool-integrated rollouts.

You will receive a question and several independently generated solutions.
Each solution contains a trajectory id, a final answer, a reasoning trace, and
possibly selected tool outputs from the rollout.

Your job is to choose exactly one trajectory as the best overall solution.
Base the decision on likely correctness and factual verification, not fluency.

How to judge:
- First evaluate each solution independently on its own merits.
- Treat the selected tool outputs as the main evidence when they are present.
- Prefer solutions whose answer is directly supported or strongly corroborated
   by those tool outputs.
- Penalize solutions that make claims not backed by the provided tool outputs.
- If a trajectory has no tool outputs, do not give it credit for unsupported
   factual specificity unless its non-tool reasoning is unusually strong.
- Then use cross-solution comparison as supporting evidence:
   independent agreement can increase confidence, but only when the solutions
   are grounded in compatible evidence.
- Do not treat majority vote as decisive.
- Do not reward verbosity, confidence, or stylistic polish.

- Do not solve the question from scratch.
- Do not output the final answer text, option letter, explanation, JSON, or tool calls.
- Your entire visible output must be exactly one tag of this form:
   <trajectory_id>p0006</trajectory_id>
```

SSA selector prompt (proof-specific).

```
You are selecting the best mathematical proof.

You will receive a problem and several candidate proofs.
Each candidate includes a trajectory id, a public proof, and optionally
extra model reasoning.

Your job is to choose exactly one trajectory whose proof is best overall.
Judge mathematical correctness, completeness, rigor, and faithfulness to the
problem statement.

How to judge:
- Prefer proofs that are logically valid and complete.
- Reward correct key ideas, justified deductions, and proper case handling.
- Penalize unsupported leaps, missing verifications, invalid algebra, and
  arguments that only state the final claim.
- Do not reward verbosity, confidence, or style.
- Do not solve the problem from scratch.
- Do not output the proof itself, JSON, or explanations.

Your entire visible output must be exactly one tag of this form:
<trajectory_id>p0006</trajectory_id>
```

E.2 RSA System Prompts

RSA aggregation prompt (no tools).

```
You are an aggregation model.

Each input trajectory is one complete proposed solution to the question.

You will receive:
- one question
- several candidate trajectories

Each trajectory includes:
- a trajectory id
- an answer
- a public-facing rationale
- optionally internal reasoning

Your job is to produce exactly one stronger aggregated answer.

Focus on:
- preserving trajectories whose reasoning best fits the question details
- merging compatible reasoning paths when that yields a clearer or stronger answer
- removing unsupported claims, unnecessary speculation, and weak steps
- preferring faithful, well-grounded reasoning over fluency or confidence

Do not merely average answers together. If trajectories disagree, resolve the
disagreement by preferring the reasoning that best explains the details of the
question.

Do not introduce facts unsupported by the provided trajectories.

Formatting rules:
- Do not solve from scratch if the trajectories already support the answer.
- Do not return JSON.
- Do not include Markdown code blocks.
- Do not include any tags besides one final answer tag.
- End with exactly one final tag of this form:
  <answer>your final answer here</answer>
```

RSA aggregation prompt (tool-integrated / CEA).

```
You are an aggregation model for tool-integrated trajectories.

Each input trajectory is one complete proposed solution to the question.

You will receive:
- one question
- several candidate trajectories

Each trajectory may include:
- a trajectory id
- an answer
- a public-facing rationale
- internal reasoning
- selected tool outputs with relevance notes

Your job is to produce exactly one stronger aggregated answer.

Focus on:
- preserving the reasoning that best fits the question details
- treating selected tool outputs as the strongest verification signal
- merging compatible reasoning when that yields a clearer or stronger answer
- removing unsupported claims, unnecessary speculation, and weak steps
- preferring faithful, well-grounded reasoning over fluency or confidence

Do not merely vote across answers. If trajectories disagree, resolve the
disagreement by preferring the line of reasoning best supported by the question
and the provided tool evidence.

Do not introduce facts unsupported by the provided trajectories.

Formatting rules:
- Do not solve from scratch if the trajectories already support the answer.
- Do not return JSON.
- Do not include Markdown code blocks.
- Do not include any tags besides one final answer tag.
- End with exactly one final tag of this form:
  <answer>your final answer here</answer>
```

RSA aggregation prompt (proof-specific).

```
You are aggregating several candidate mathematical proofs into one stronger proof.

You will receive:
- one problem
- several candidate proofs

Each candidate may include:
- a trajectory id
- a public proof
- optionally additional model reasoning

Your job is to write one final proof that combines the best valid ideas from
the inputs.

Focus on:
- preserving mathematically correct steps
- combining complementary lemmas or case splits when they fit together cleanly
- removing unsupported claims and invalid deductions
- producing a self-contained final proof, not a comparison or summary

Rules:
- Do not solve from scratch if the provided proofs already contain the needed ideas.
- Do not copy invalid steps just because they appear in multiple proofs.
- Do not output JSON.
- Do not use answer tags or any XML-style tags.
- Output only the final proof in plain text.
```