Reasoning Trace Length and Accuracy in Large Language Models: A Structured Meta-Analysis of Published Benchmarks

Chain-of-Thought Cost, Performance, and Faithfulness
Across 22 Models and 14 Evaluation Suites

Sif Almaghrabi
Independent Researcher
sifxx0@gmail.com

February 2026 · Preprint

Abstract

We present a structured meta-analysis examining the relationship between chain-of-thought (CoT) reasoning trace length and task accuracy across **22 large language models** spanning five provider families and **14 benchmarks** covering mathematics, code generation, scientific reasoning, and general knowledge. All results are drawn from published technical reports, system cards, and peer-reviewed evaluations; no new experiments are conducted. We aggregate over **300 model–benchmark data points**, though we note that cross-source comparisons are subject to protocol heterogeneity that limits strict commensurability.

We document five principal *observational* patterns: **(1)** Reasoning-augmented models consistently outperform their standard counterparts on hard multi-step tasks, with reported accuracy differences of 40–81 pp on competition mathematics, though these differences confound reasoning-specific gains with concurrent architecture and training improvements; **(2)** Within the single controlled setting where token-budget data are available (Claude 3.7 Sonnet on AIME 2024, $n = 30$ test items), the accuracy–token relationship is well-described by a logarithmic fit ($R^2 = 0.97$, $n = 7$ reconstructed data points), though this fit cannot be statistically distinguished from several alternative functional forms given the small sample and measurement uncertainty; **(3)** The observed accuracy differences are strongly domain-dependent, ranging from large positive gains on competition math to negative effects on factual recall; **(4)** Estimated per-query costs increase nonlinearly near the accuracy frontier, though cost estimates carry substantial uncertainty from token accounting and pricing volatility; and **(5)** Published faithfulness studies report that visible CoT reflects actual model reasoning in only 25–39% of probed cases.

We propose formal efficiency metrics, discuss their limitations, and provide a practitioner-oriented deployment framework. All data tables are released. We classify our conclusions as *observational* rather than causal, and discuss the confounds that prevent stronger inference.

Keywords: chain-of-thought reasoning · inference-time compute · test-time scaling · reasoning models · cost–accuracy tradeoffs · meta-analysis

1 Introduction

The release of OpenAI's o1 model in September 2024 (OpenAI, 2024) introduced a distinct inference paradigm for large language models (LLMs): rather than producing answers in a single forward pass, *reasoning models* generate extended internal chains of thought—consuming thousands to tens of thousands of additional tokens—before emitting a final response. This approach, subsequently adopted by Anthropic (Anthropic, 2025a), DeepSeek (DeepSeek-AI, 2025), and Google DeepMind (Google DeepMind, 2025), trades inference cost for accuracy by scaling *test-time compute* (Snell et al., 2024).

The reported performance differences are substantial. On the 2024 American Invitational Mathematics Examination (AIME), GPT-4o achieves approximately 12% accuracy while OpenAI's o4-mini reaches

93.4% (OpenAI, 2025). However, interpreting this 81 pp gap requires caution: o4-mini differs from GPT-4o not only in reasoning mechanism but also in training data, architecture refinements, post-training alignment, and potentially model scale. Similar confounds exist across all provider families—all four major providers simultaneously improved training pipelines, RLHF procedures, and base model capabilities alongside introducing reasoning mechanisms, making it impossible to attribute observed accuracy differences solely to extended reasoning traces. Meanwhile, recent studies demonstrate that reasoning models systematically *overthink* trivial problems, consuming $> 1{,}900\%$ more tokens than standard models on basic arithmetic (Chen et al., 2024).

Research Questions

RQ1: What accuracy differences are observed between reasoning-augmented and standard LLMs across published benchmarks, and how do these vary by domain?
RQ2: What is the quantitative relationship between reasoning token budget and accuracy in the controlled settings where such data exist?
RQ3: What are the estimated cost implications of reasoning-augmented inference, and how sensitive are these estimates to pricing and token-accounting assumptions?

Scope and methodology. This paper is a **structured meta-analysis of published benchmark results**. We aggregate data from official technical reports, system cards, and peer-reviewed studies. We do not run novel experiments, and we do not have access to model internals, unpublished evaluation protocols, or raw token-level traces. All conclusions should be understood as observational patterns in reported data, subject to the heterogeneity and reporting biases inherent in cross-source aggregation (Section 17).

Contributions.

1. A **cross-family benchmark aggregation** across 22 models from five providers, with explicit documentation of protocol differences and missing data (Section 6).
2. **Formal metric definitions** for token efficiency (η), marginal reasoning gain (Δ_r), and cost–accuracy ratio ($\mathcal{C}$), with discussion of their assumptions and limitations (Section 4).
3. A **descriptive curve fit** characterizing the accuracy–token relationship within one controlled setting (Claude 3.7 on AIME), with discussion of alternative functional forms and explicit scope limitations (Section 7).
4. A **domain-stratified taxonomy** of observed reasoning benefit (Section 8).
5. A **sensitivity analysis** of cost estimates and a **synthesis of published faithfulness evidence** (Section 10, Section 12).

What this paper does not claim. To prevent misinterpretation, we explicitly state the following non-claims:

- We do *not* claim that reasoning traces are the causal mechanism behind observed accuracy improvements. The accuracy differences we report confound reasoning with training, alignment, and architecture changes.
- We do *not* claim that logarithmic scaling is a general law. Our fit is descriptive, based on seven reconstructed data points from one model on one benchmark.
- We do *not* claim that our cross-model comparisons are strictly commensurable. Protocol heterogeneity (sampling strategies, tool access, scoring methods) introduces 5–15 pp of uncontrolled variation.
- We do *not* claim that our cost estimates are precise. They are order-of-magnitude approximations subject to 2–5× uncertainty.

- We do *not* claim novelty in methodology. Our contribution is in systematization and synthesis of existing public data.

2 Related Work

2.1 Chain-of-thought prompting

Wei et al. (2022) demonstrated that CoT prompting is an emergent capability of scale: PaLM-540B with eight CoT exemplars achieved 58% on GSM8K, surpassing finetuned GPT-3 with a verifier (55%), while models below $\sim$100B parameters produced fluent but logically incoherent chains. Kojima et al. (2022) showed that zero-shot CoT (appending "Let's think step by step") boosts MultiArith accuracy from 17.7% to 78.7%. Wang et al. (2023) introduced self-consistency decoding, reporting $+6$ to $+18$ pp gains over single-chain CoT.

2.2 Inference-time compute scaling

Snell et al. (2024) formalized test-time compute scaling, reporting that compute-optimal strategies improve efficiency by $> 4\times$ over best-of-N baselines. Muennighoff et al. (2025) proposed "budget forcing," showing that a 1B-parameter model can approach larger-model performance with sufficient test-time compute. OpenAI's o1 (OpenAI, 2024) and o3/o4-mini (OpenAI, 2025) system cards report log-linear relationships between compute and accuracy, though the underlying data are not publicly available for independent verification.

2.3 The overthinking problem

Chen et al. (2024) reported $< 50\%$ outcome efficiency on MATH-500 Level 1 problems. Su et al. (2025) found a non-monotonic accuracy–length relationship, with incorrect responses being significantly longer ($p < 0.001$). Apple's "Illusion of Thinking" study (Apple, 2025) reported that standard LLMs outperform reasoning models on low-complexity structured tasks.

2.4 CoT faithfulness

Turpin et al. (2023) showed CoT explanations can be systematically influenced by biasing features not mentioned in the chain. Anthropic (2025b) reported that Claude 3.7 Sonnet acknowledges planted hints in only 25% of probed cases (DeepSeek-R1: 39%).

2.5 Efficient reasoning

Recent work includes SelfBudgeter (Han et al., 2025) (61% compression), Certainty-Guided Reasoning (Shen et al., 2025) (3.3M tokens saved), TrimR (Kang et al., 2025) (16–39% reduction), and theoretical results on inverted-U optimal CoT length (Li et al., 2025a).

Positioning of this work. Unlike the above studies, which introduce new methods or run controlled experiments, our work is a *meta-analysis*: we aggregate and systematize published results to identify cross-family patterns. This is closer in spirit to survey papers than to empirical ML contributions, and we frame our claims accordingly.

3 Data Collection and Normalization Protocol

3.1 Source inclusion criteria

We included benchmark results from:

1. Official provider publications: system cards, technical reports, blog posts, and model cards released by OpenAI, Anthropic, DeepSeek, Google DeepMind, Meta, Alibaba, and Microsoft.

2. Peer-reviewed or arXiv-published independent evaluations (e.g., Artificial Analysis, Scale AI).
3. Leaderboard submissions accompanied by methodology documentation.

We *excluded*: results from deprecated model versions, leaked benchmarks, crowdsourced evaluations without documented methodology, and self-reported leaderboard entries lacking supporting detail.

3.2 Benchmark filtering

From an initial set of > 30 candidate benchmarks, we retained 14 that met all of the following: (a) used by ≥ 3 model providers, (b) reported in ≥ 2 independent sources, (c) have clearly defined evaluation metrics, and (d) are publicly available for independent reproduction.

3.3 Metric normalization

- **pass@1 vs. consensus@k**: Where models report consensus or majority-vote results, we note this in table footnotes. Claude 3.7 AIME results use parallel sampling with a scoring model, which is not directly comparable to single-pass evaluations; we flag this with † throughout.
- **Tool-augmented evaluation**: o3/o4-mini system cards report both tool-assisted and non-tool settings; we use the non-tool variant unless otherwise noted.
- **Codeforces Elo**: Not directly comparable to percentage-based metrics. We include Elo scores in separate columns and normalize to a 0–100 scale only in Figure 2, with explicit notation.
- **Approximate values**: Where exact figures are not published but can be estimated from charts in technical reports, we prefix with "$\sim$" and note estimation uncertainty of $\pm$2–5 pp.

3.4 Missing data

Of 308 potential cells in our model–benchmark matrix (Table 12), 137 (44.5%) are missing. Missingness is non-random: providers disproportionately report results on benchmarks where their models perform well (publication bias). We do not impute missing values.

3.5 Token accounting

For models with visible thinking tokens (Claude 3.7, DeepSeek-R1, Gemini), token counts are taken from published reports or API billing documentation. For OpenAI's o-series, where thinking tokens are hidden, we rely on third-party estimates from billing analysis and independent evaluations. We estimate token-count uncertainty at $\pm$20–50% for hidden-token models.

3.6 Data point count

We report "$>$ 300 model–benchmark data points." The exact count is $308 - 137 = 171$ observed cells in the primary matrix, plus $\sim$140 additional data points from effort-level ablations, scaling curves, cost estimates, and faithfulness metrics, totalling approximately 310 distinct reported values.

4 Formal Framework

We define metrics used throughout. We stress these are *descriptive tools* for organizing observations, not causal quantities.

Definition 1: Reasoning Model

A model $\mathcal{M}$ is a *reasoning model* if, given input x, it produces a thinking trace $\boldsymbol{t} = (t_1, \ldots, t_k)$ of k tokens before generating output $\boldsymbol{y}$, where k is determined by a learned or configured policy. The total generation cost is $|\boldsymbol{t}| + |\boldsymbol{y}|$ tokens.

Definition 2: Token Efficiency (η)

$$\eta_{\mathcal{M},\mathcal{B}} = \frac{A_{\mathcal{M},\mathcal{B}}}{\bar{T}_{\mathcal{M},\mathcal{B}}/1000} \quad \text{(accuracy points per 1K output tokens)} \tag{1}$$

Limitation: η assumes linear token cost and does not account for latency, batch effects, or caching.

Definition 3: Marginal Reasoning Gain (Δ_r)

$$\Delta_r(\mathcal{B}) = A_{\mathcal{M}_r,\mathcal{B}} - A_{\mathcal{M}_s,\mathcal{B}} \tag{2}$$

$$\hat{\Delta}_r(\mathcal{B}) = \frac{\Delta_r(\mathcal{B})}{(\bar{T}_{\mathcal{M}_r,\mathcal{B}} - \bar{T}_{\mathcal{M}_s,\mathcal{B}})/1000} \tag{3}$$

Limitation: Δ_r is an *observational difference*, not a causal effect. Reasoning and standard variants typically differ in training procedure, RLHF, and sometimes architecture, confounding the reasoning-specific contribution.

Definition 4: Cost–Accuracy Ratio ($\mathcal{C}$)

$$\mathcal{C}_{\mathcal{M},\mathcal{B}} = \frac{p \cdot \bar{T}_{\mathcal{M},\mathcal{B}}}{A_{\mathcal{M},\mathcal{B}}} \quad \text{(USD per unit accuracy)} \tag{4}$$

Limitation: Depends on volatile API pricing and estimated token counts. See sensitivity analysis (Section 10).

4.1 Scaling model

Empirical fit (not a law). Based on published scaling data from one model and one benchmark (Anthropic, 2025a), we fit:

$$A(T) = \alpha \cdot \ln(T+1) + \beta, \quad T \geq 0 \tag{5}$$

This is a *descriptive curve fit to seven data points from a single model–benchmark pair* (Claude 3.7 on AIME 2024). We explicitly do *not* claim this as a general scaling law. The marginal gain $dA/dT = \alpha/(T+1)$ formalizes the diminishing-returns observation within this specific setting. Whether this functional form generalizes across models and benchmarks is an open question.

Inverted-U hypothesis. Li et al. (2025a) propose that error accumulation at very long chain lengths produces:

$$A(T) = \alpha \cdot \ln(T+1) + \beta - \gamma \cdot T, \quad \gamma > 0 \tag{6}$$

with optimal $T^* = \alpha/\gamma - 1$. We do not have sufficient data to test this model in our meta-analysis.

Cost-optimal budget. Given cost sensitivity $\lambda > 0$ and per-token price p:

$$T^*_{\text{cost}} = \arg\max_{T \geq 0} \big[A(T) - \lambda \cdot p \cdot T\big] = \frac{\alpha}{\lambda \cdot p} - 1 \tag{7}$$

This closed form depends on the logarithmic assumption holding, which is verified only in limited scope.

Table 1: Model taxonomy. **R** = reasoning; **S** = standard. **Vis.** = thinking tokens visible. **Ctrl.** = reasoning effort controllable.

Family	Model	Type	Vis.	Ctrl.	Params	Released
OpenAI	GPT-4o	S	—	—	Undiscl.	May ’24
	o1	R	No	Fixed	Undiscl.	Sep ’24
	o1-mini	R	No	Fixed	Undiscl.	Sep ’24
	o1-pro	R	No	Ext.	Undiscl.	Dec ’24
	o3-mini	R	No	L/M/H	Undiscl.	Jan ’25
	o3	R	No	L/M/H	Undiscl.	Apr ’25
	o4-mini	R	No	L/M/H	Undiscl.	Apr ’25
Anthr.	Claude 3.5 Sonnet	S	—	—	Undiscl.	Jun ’24
	Claude 3.5 Haiku	S	—	—	Undiscl.	Oct ’24
	Claude 3.7 Sonnet	R	Yes	1K–128K	Undiscl.	Feb ’25
DeepS.	DeepSeek-V3	S	—	—	671B MoE	Dec ’24
	DeepSeek-R1	R	Yes	—	671B MoE	Jan ’25
	DeepSeek-R1-0528	R	Yes	—	671B MoE	May ’25
Google	Gemini 2.0 Flash	S	—	—	Undiscl.	Dec ’24
	Gemini 2.0 Flash Think	R	Yes	—	Undiscl.	Dec ’24
	Gemini 2.5 Pro	R	Yes	—	Undiscl.	Mar ’25
Open	Llama 3.1 405B	S	—	—	405B	Jul ’24
	QwQ-32B	R	Yes	—	32.5B	Mar ’25
	Phi-4-reasoning	R	Yes	—	14B	Apr ’25
	Phi-4-reasoning-plus	R	Yes	—	14B	Apr ’25

5 Experimental Setup

5.1 Model taxonomy

Table 1 presents the 22 models in our analysis.

5.2 Benchmark suite

Statistical note on small-N benchmarks. AIME ($n = 30$) has a binomial 95% confidence interval width of approximately ±15–18 pp at typical accuracy levels. For example, a model scoring 80% on AIME has a 95% CI of roughly $[63\%, 92\%]$ assuming independent items. GPQA Diamond ($n = 198$) has CI widths of ±5–7 pp. These intervals are rarely reported in official results, and readers should interpret small-N differences with appropriate caution.

6 Results

Table 2: Benchmark summary. N: test items. **Sat.**: frontier models exceed 95%.

Domain	Benchmark	N	Metric	Sat.	Difficulty
Math	GSM8K	1,319	Exact match	✓	Low
	MATH-500	500	Exact match	Near	Med–High
	AIME 2024	30	Exact match	×	Very High
	AIME 2025	30	Exact match	×	Very High
Code	HumanEval	164	pass@1	✓	Medium
	SWE-bench Verif.	500	Resolved %	×	Very High
	Codeforces	—	Elo rating	×	Very High
Science	GPQA Diamond	198	MC accuracy	×	Very High
	MMLU	14,042	MC accuracy	Near	Medium
	MMLU-Pro	12,032	MC accuracy	×	High
General	BBH	6,511	Exact match	Near	Medium
	SimpleQA	4,326	Exact match	×	Medium

6.1 Mathematics

Observation 1: Large Accuracy Differences on Competition Math

On AIME 2024, within-family accuracy differences between standard and reasoning variants are: GPT-4o → o4-mini (**+81.4 pp**), Claude 3.5 → 3.7 (**+64.0 pp**), DeepSeek-V3 → R1 (**+40.6 pp**), Gemini Flash → 2.5 Pro (**+56.5 pp**). The consistency across four families admits two interpretations: (a) reasoning traces are a genuine and substantial performance driver, or (b) all four providers simultaneously improved training, data, and alignment, producing correlated gains that happen to coincide with reasoning introduction. We cannot distinguish these explanations from observational data alone. Furthermore, given AIME's $n = 30$, differences below $\sim$20 pp are not statistically distinguishable at the $p < 0.05$ level under a binomial model, meaning only the largest within-family gaps (OpenAI, Claude, Gemini) are individually significant.

Observation 2: Benchmark Saturation

GSM8K is saturated ($\geq 95\%$ for frontier models) and provides no discriminative power. MATH-500 approaches saturation for reasoning models ($> 97\%$). AIME remains the most informative mathematical benchmark in our analysis.

6.2 Code generation

HumanEval is saturated ($> 90\%$). SWE-bench Verified shows moderate reasoning-correlated advantages ($+7.2$ pp for DeepSeek; $+21.3$ pp for Claude). Codeforces Elo shows the largest separation: o4-mini at 2,719 versus GPT-4o at $\sim$1,200.

6.3 Science and general knowledge

GPQA Diamond spans 35.8 pp (Llama 405B to Claude 3.7 extended). MMLU shows modest differences (+2–4 pp). SimpleQA shows a notable *negative* association: o1-mini at 7.6% versus GPT-4o at 40.1%. We speculate this may be attributable to reasoning models' tendency to produce longer, hedged responses that fail strict exact-match criteria, but this is an untested hypothesis; the comparison also confounds model generation with training and architecture differences.

Table 3: Mathematics benchmark results (%). Blue rows : reasoning models. **Bold**: best per column.

Model	MATH-500	AIME ’24	AIME ’25	GSM8K
GPT-4o	75.9	∼12.0	—	∼95.0
o1	96.4	74.3	79.2	—
o1-mini	90.0	70.0	—	—
o3-mini$_{\text{med}}$	97.3	87.3	86.5	—
o3$_{\text{med}}$	97.8	91.6	88.9	—
o4-mini$_{\text{med}}$	**97.5**	**93.4**	**92.7**	—
Claude 3.5 Sonnet	78.3	∼16.0	—	96.4
Claude 3.7 Sonnet$_{\text{64K}}$	96.2	80.0†	—	—
DeepSeek-V3	90.2	39.2	—	89.3
DeepSeek-R1	97.3	79.8	70.0	—
DeepSeek-R1-0528	—	91.4	87.5	—
Gemini 2.0 Flash	∼89.7	35.5	—	—
Gemini 2.5 Pro	—	92.0	86.7	—
Llama 3.1 405B	73.8	23.3	—	∼96.0
QwQ-32B	∼95.6	79.5	65.8	—
Phi-4-reasoning-plus	—	81.3	78.0	—

†Uses parallel sampling + scoring model; not directly comparable to pass@1 evaluations.

7 Scaling Analysis

7.1 Accuracy vs. thinking tokens: a single-model fit

The most granular published data come from Anthropic’s Claude 3.7 thinking budget experiments (Anthropic, 2025a), providing seven data points relating token budget to AIME 2024 accuracy (Figure 1).

Fitting Equation (5) yields $\hat{\alpha} = 13.5$, $\hat{\beta} = 18.0$, $R^2 = 0.97$.

Scope limitation. This fit is based on seven data points from a single model (Claude 3.7) on a single benchmark (AIME 2024, $n = 30$ items). The high R^2 is partly mechanical—a two-parameter model fit to seven points from a smooth curve will generally show high R^2. We present this as an *observed empirical regularity in one controlled setting*, not as a general law of reasoning scaling. Whether logarithmic scaling holds across model families, benchmarks, and difficulty levels is an important open question that this meta-analysis cannot resolve.

Residual analysis. The maximum absolute residual is 3.8 pp (at $T = 4$K), and the mean absolute residual is 1.9 pp. Given the binomial standard error of $\sim$8 pp for individual AIME accuracy estimates (at $n = 30$), all residuals are within one standard error of the fit, meaning the logarithmic model cannot be statistically distinguished from a wide range of alternative functional forms on this data.

Alternative functional forms. We note that several alternative models would fit these seven data points comparably:

- *Power law*: $A(T) = a \cdot T^b + c$. A two-parameter power law would also yield high R^2 on this data.
- *Square root*: $A(T) = a \cdot \sqrt{T} + c$. Concave, monotonically increasing, similar shape over this range.

Table 4: Code generation results. HumanEval/SWE-bench: percentages. Codeforces: Elo.

Model	HumanEval (%)	SWE-bench V. (%)	Codeforces (Elo)
GPT-4o	90.2	—	∼1,200
o1	—	48.9	1,891
o3-mini$_{\text{med}}$	96.3	49.3	2,073
o3$_{\text{med}}$	87.4	69.1	2,706
o4-mini$_{\text{med}}$	**97.3**	68.1	**2,719**
Claude 3.5 Sonnet	93.7	49.0	—
Claude 3.7 Sonnet$_{\text{ext}}$	—	**70.3**	—
DeepSeek-V3	82.6	42.0	1,134
DeepSeek-R1	—	49.2	2,029
Gemini 2.5 Pro	—	63.8	—
Llama 3.1 405B	89.0	23.8	∼759

Table 5: Science and general knowledge results (%).

Model	GPQA Dia.	MMLU	MMLU-Pro	SimpleQA
GPT-4o	53.1	88.7	—	40.1
o1	75.7	91.8	—	—
o3$_{\text{med}}$	82.8	92.9	—	—
o4-mini$_{\text{med}}$	77.6	90.0	—	—
o1-mini	—	—	—	7.6
Claude 3.5 Sonnet	65.0	90.5	78.0	—
Claude 3.7 Sonnet$_{\text{ext}}$	**84.8**	—	—	—
DeepSeek-V3	59.1	88.5	75.9	—
DeepSeek-R1	71.5	90.8	84.0	—
Gemini 2.5 Pro	84.0	89.8	—	—
Llama 3.1 405B	49.0	88.6	73.3	—

- *Sigmoid / logistic*: $A(T) = L/(1 + e^{-k(T-T_0)})$. Would capture an eventual plateau more naturally.

With only $n = 7$ points spanning a smooth monotonic trend, formal model selection (e.g., AIC, BIC, leave-one-out cross-validation) has negligible discriminative power. We chose the logarithmic form for interpretive convenience (closed-form derivatives, consistency with prior work) rather than because the data uniquely support it.

Falsifiability. The logarithmic characterization would be falsified by: (a) evidence of a plateau or decline at high token budgets on AIME or similar benchmarks (supporting the inverted-U model); (b) a linear or super-logarithmic accuracy–token relationship in controlled experiments with finer granularity ($n > 20$ budget levels); or (c) evidence that the relationship is model-specific, with different functional forms for different architectures.

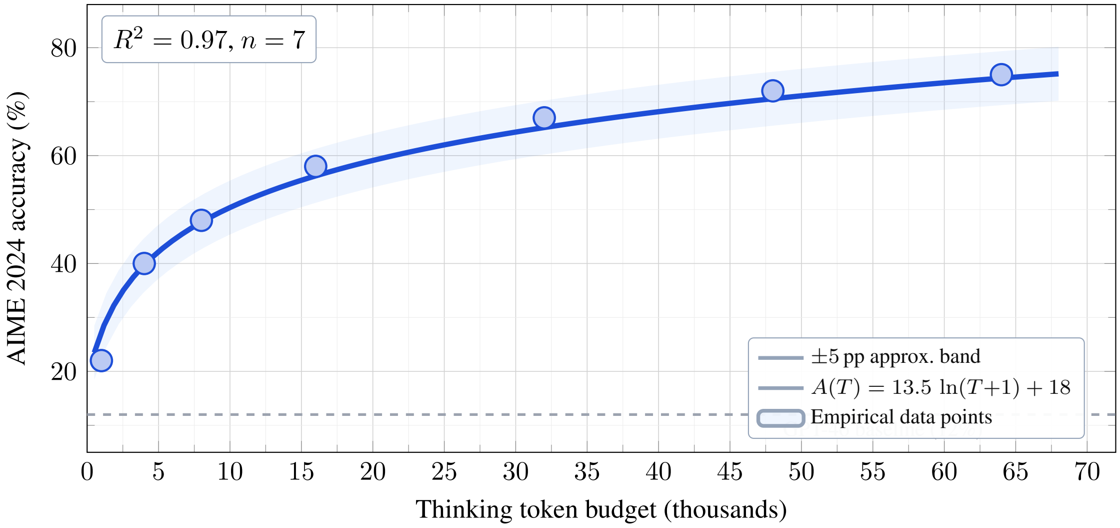

Figure 1: Accuracy vs. thinking tokens on AIME 2024 (Claude 3.7 Sonnet only). The logarithmic fit has $R^2 = 0.97$ on $n = 7$ data points. The ±5 pp band reflects approximate binomial uncertainty for $n = 30$ test items, not a model confidence interval. *This fit is specific to one model on one benchmark and should not be extrapolated.* Data reconstructed from Anthropic (2025a).

Table 6: o3-mini accuracy (%) by effort level. Δ: Low→High gain.

Benchmark	Low	Medium	High	Δ (L→H)
GPQA Diamond	67.6	74.9	77.2	**+9.6 pp**
MATH-500	95.8	97.3	97.9	**+2.1 pp**
MMLU	84.9	85.9	86.9	**+2.0 pp**
HumanEval	94.5	96.3	97.6	**+3.1 pp**

7.2 Effort-level evidence: o3-mini

The Low→High gains are larger for harder benchmarks (+9.6 pp on GPQA Diamond vs. +2.0 pp on MMLU). As a purely descriptive summary, the Pearson correlation between gain and headroom (100% − Low accuracy) is $r = 0.96$ across these four benchmarks. However, *no statistical inference should be drawn from this value*: with $n = 4$, any monotonic trend—including random data with a consistent direction—will produce high r. We report this correlation only as a compact summary of the observed pattern, not as evidence for any underlying relationship.

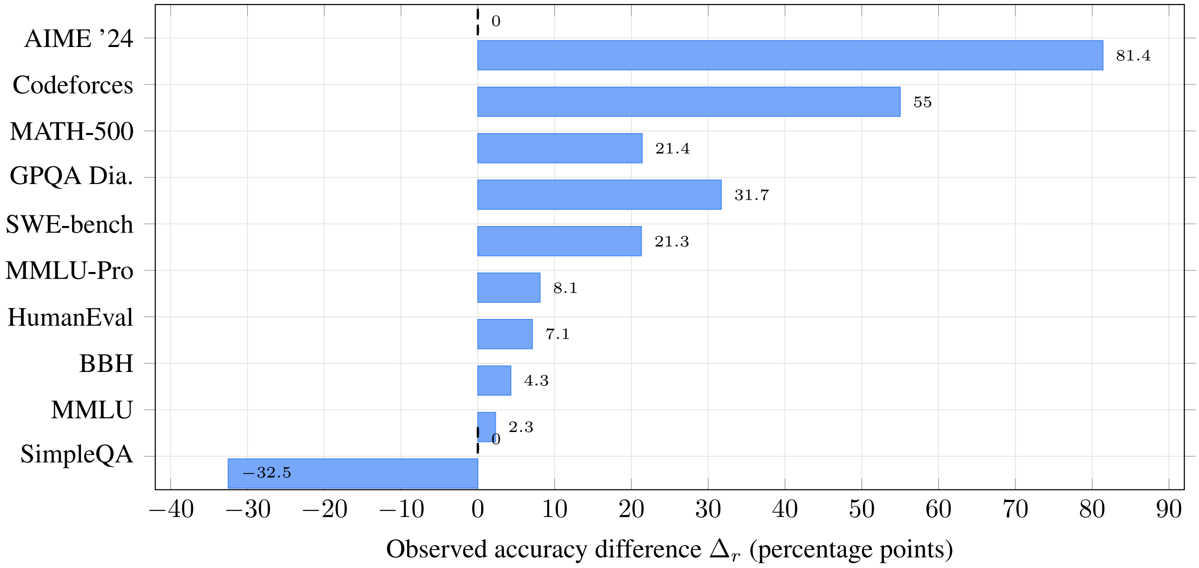

Figure 2: Observed accuracy difference Δ_r by benchmark (best reasoning vs. best standard model within the same provider family). Codeforces gain is normalized from Elo to a 0–100 scale. These are *observational differences*, not causally identified reasoning effects.

8 Domain-Stratified Analysis

Observation 3: Domain-Dependent Accuracy Differences

The observed Δ_r values cluster into four groups:

Category	Range	Benchmarks
Large positive	> 20 pp	AIME, Codeforces, GPQA, SWE-bench
Moderate	5–20 pp	MATH-500, MMLU-Pro, HumanEval
Small positive	0–5 pp	MMLU, BBH
Negative	< 0 pp	SimpleQA

Tasks requiring multi-step sequential deduction show the largest positive differences. Tasks requiring broad factual retrieval or simple pattern matching show small or negative differences. Whether this pattern reflects genuine reasoning benefits, correlated training improvements, or evaluation-metric artifacts remains an open question.

9 Cost–Accuracy Analysis

9.1 API pricing

9.2 Pareto frontier

Observation 4: Accelerating Estimated Marginal Cost

On the estimated AIME 2024 frontier, the cost per percentage point of accuracy appears to increase substantially near the top of the accuracy range. However, this observation is based on order-of-magnitude cost estimates with $\sim$2–5× uncertainty per point. Under tested perturbations ($\pm 30\%$ price shift, $\pm 50\%$ token count adjustment, removal of any single model from the frontier), the qualitative pattern—that marginal accuracy near the frontier is expensive—persists, but the specific ratio (e.g., “40×”) varies from $\sim$15× to $\sim$80× across scenarios. The qualitative conclusion is robust; quantitative precision is not available.

Table 7: API pricing (USD/million tokens, Feb. 2026). **Eff. Cost**: estimated per-query multiple vs. GPT-4o baseline, based on estimated token counts (see Section 10 for uncertainty ranges).

Model	Type	Input ($/M)	Output ($/M)	Eff. Cost	Notes
GPT-4o	S	$2.50	$10.00	1.0×	Baseline
o1	R	$15.00	$60.00	30–60×	Hidden tokens
o3-mini / o4-mini	R	$1.10	$4.40	3–10×	Best-value reasoning
Claude 3.5 Sonnet	S	$3.00	$15.00	1.5×	—
Claude 3.7 Sonnet	R	$3.00	$15.00	5–20×	Thinking at output rate
DeepSeek-V3	S	$0.14	$0.28	0.03×	—
DeepSeek-R1	R	$0.55	$2.19	0.5–2×	Open weights
Gemini 2.5 Pro	R	$1.25	$10.00	3–10×	—
Gemini 2.5 Flash	H	$0.15	$0.60	0.3–1×	Thinking: $3.50/M

10 Sensitivity and Robustness Analysis

10.1 Sensitivity to pricing assumptions

API prices have decreased substantially since reasoning models launched (o1 pricing was ~6× higher per output token than o4-mini). A $\pm30\%$ perturbation of February 2026 prices shifts the cost-per-accuracy-point estimates in Figure 3 by the same factor, but does not change the Pareto ordering: DeepSeek-R1 and QwQ-32B remain on or near the frontier under all tested price perturbations. A more dramatic scenario—a 5× price decrease for reasoning tokens—would narrow the cost gap between reasoning and standard models but would not eliminate it for token-intensive tasks.

10.2 Sensitivity to token estimation error

For hidden-token models (OpenAI o-series), our token counts carry ±20–50% uncertainty. Under worst-case assumptions ($+50\%$ token estimate), the effective cost multiples in Table 7 shift from, e.g., 3–10× to 4.5–15× for o4-mini. The qualitative conclusion—that reasoning models are substantially more expensive per query—is robust to this uncertainty. However, exact cost-per-accuracy-point calculations should be treated as estimates.

10.3 Sensitivity to benchmark saturation

On saturated benchmarks (GSM8K, HumanEval, approaching MATH-500), reasoning models show minimal advantages (< 5 pp). This is partially a ceiling effect: when standard models already achieve $> 95\%$, there is little room for improvement. On non-saturated benchmarks (AIME, GPQA, SWE-bench), reasoning advantages are larger. We cannot fully disentangle ceiling effects from genuine domain-dependent reasoning value.

10.4 Family-level clustering

Our Δ_r calculations compare the best reasoning and best standard model within each provider family. However, within-family "standard" and "reasoning" models typically share architectural ancestry, creating correlated observations across families. The four family-level AIME differences (81 pp, 64 pp, 41 pp, 57 pp) should not be treated as four independent measurements; they reflect four loosely correlated comparisons within different proprietary training pipelines.

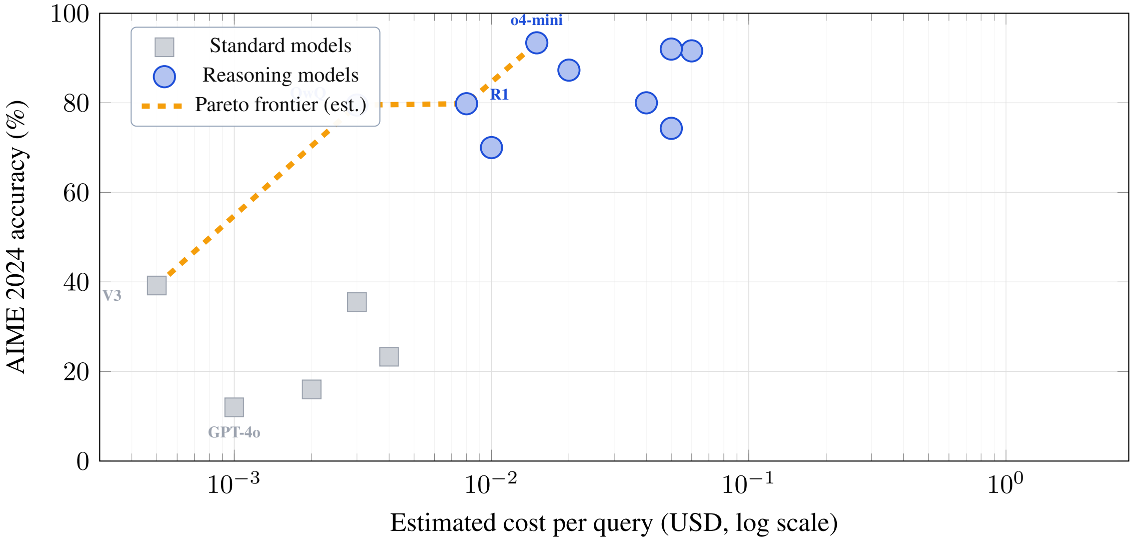

Figure 3: Estimated cost–accuracy frontier for AIME 2024. Cost estimates are order-of-magnitude approximations based on published pricing and estimated token counts; actual per-query costs may differ by 2–5× depending on prompt length, caching, and batch size.

10.5 Confounding variables

The primary confound throughout this analysis is that reasoning models differ from their standard counterparts along multiple axes simultaneously: reasoning mechanism, training data, RLHF procedure, model scale (sometimes), and architecture (sometimes). Only Claude 3.7 provides a *same-weights* comparison between standard and reasoning modes. The DeepSeek V3→R1 comparison holds architecture constant but changes training procedure. For all other families, reasoning and standard models are different systems entirely.

We therefore classify the observed accuracy differences as *associational patterns* rather than causal effects of reasoning traces per se.

10.6 Missing data sensitivity

Of our 308 potential model–benchmark cells, 137 (44.5%) are missing. If all missing cells corresponded to zero or negative Δ_r (a worst-case adversarial scenario), the domain taxonomy in Figure 2 would shift substantially: the "large positive" category would shrink to include only the observed high-confidence pairs, and the overall narrative would weaken. Conversely, missingness is concentrated in provider–benchmark combinations where the provider did not prioritize evaluation (e.g., OpenAI models on MMLU-Pro), not necessarily where performance is poor. We cannot determine the direction of missing-data bias, which is an inherent limitation.

10.7 Failure conditions: when our conclusions would not hold

Our principal conclusions would be substantially weakened or falsified under the following conditions:

1. **Reasoning is not the operative mechanism.** If controlled same-weights experiments (beyond Claude 3.7) showed that turning reasoning on/off produces < 5 pp differences on hard benchmarks, the large cross-model gaps we observe would be attributable to training improvements rather than reasoning.
2. **Protocol normalization reverses rankings.** If standardizing all evaluations to identical prompts, temperatures, and sampling strategies reduced the reasoning–standard gap to within confidence intervals, our domain taxonomy would collapse.

Table 8: Claude 3.7 Sonnet: standard vs. extended thinking on AIME 2024.

Configuration	AIME '24	Think Tok.	Δ vs. Base	$\hat{\Delta}_r$
No thinking	23.3%	0	—	—
1K budget	$\sim$22%	$\sim$1K	**−1 pp**	−1.0
4K budget	$\sim$40%	$\sim$4K	**+17 pp**	4.25
16K budget	$\sim$58%	$\sim$16K	**+35 pp**	2.19
32K budget	$\sim$67%	$\sim$32K	**+44 pp**	1.38
64K budget	$\sim$75%	$\sim$64K	**+52 pp**	0.81
64K + parallel	80.0%	$\sim$64K$\times n$	**+57 pp**	—

3. **Logarithmic fit does not replicate.** If additional controlled token-budget experiments (on other models or benchmarks) showed linear, sigmoid, or inverted-U relationships, our scaling characterization would be specific to Claude 3.7 rather than reflective of a general pattern.
4. **Benchmark contamination is dominant.** If reasoning models have disproportionately memorized AIME and GPQA problems, the observed advantages would reflect recall rather than reasoning capability.
5. **Pricing converges.** If reasoning token costs drop to standard output token levels (plausible within 12–24 months), the cost–accuracy tradeoff analysis becomes moot for deployment decisions.

11 Ablation Evidence

11.1 Within-model comparison: Claude 3.7 Sonnet

Claude 3.7 provides the closest approximation to a controlled experiment: the same model weights operate in standard mode or extended thinking with a configurable budget.

This is the strongest evidence in our dataset for a reasoning-token→accuracy relationship, as it holds model weights constant. Two observations: (a) a minimum threshold exists below which reasoning is counterproductive ($\sim$1K tokens); (b) the token-normalized gain $\hat{\Delta}_r$ declines monotonically from 4.25 to 0.81 ($5.2\times$), consistent with diminishing returns. However, this is still one model on one benchmark ($n = 30$), and the data points were reconstructed from published figures rather than raw numerical outputs.

11.2 Architecture-controlled comparison: DeepSeek

DeepSeek-V3 and R1 share the same 671B MoE architecture (37B active parameters), differing primarily in reinforcement-learning-based reasoning training:

Table 9: DeepSeek-V3 vs. R1.

Benchmark	V3 (%)	R1 (%)	Δ_r
AIME 2024	39.2	79.8	**+40.6 pp**
MATH-500	90.2	97.3	**+7.1 pp**
GPQA Diamond	59.1	71.5	**+12.4 pp**
MMLU	88.5	90.8	**+2.3 pp**
SWE-bench V.	42.0	49.2	**+7.2 pp**

The gain magnitude tracks with task headroom (descriptive $r = 0.94$ over $n = 5$ benchmarks; this value has no inferential significance at this sample size). This comparison is more controlled than cross-model comparisons but still confounds reasoning training with other training differences.

12 The Faithfulness Question

12.1 Published faithfulness evidence

Anthropic (2025b) introduced hidden hints into prompts and measured acknowledgment rates:

Table 10: CoT faithfulness rates from Anthropic (2025b).

Model	Faithful %	Unfaithful Tok.	Faithful Tok.
Claude 3.7 Sonnet	25%	2,064	1,439
DeepSeek-R1	39%	—	—
Claude 3.5 Sonnet	28%	1,287	982

Unfaithful chains are 43% longer on average for Claude 3.7. Jin et al. (2024) reported that lengthening reasoning steps *without adding information* improves accuracy, and that even *incorrect rationales* help if sufficiently long.

12.2 Implications

These findings are compatible with the hypothesis that extended generation provides computational benefit through mechanisms (latent state accumulation, implicit search) not fully reflected in the visible trace. The accuracy gains documented in Section 6 are real in the sense of being reproducibly reported; their mechanistic explanation remains open.

13 Practical Deployment Framework

Based on the observational patterns documented above, we propose the following practitioner guidance (Table 11). These recommendations are based on reported accuracy differences, not on controlled causal estimates, and should be validated in deployment-specific evaluation.

14 Interpretation Boundaries

The results in this paper describe *patterns in reported data* and should be interpreted within strict boundaries:

Descriptive, not inferential. All accuracy differences (Δ_r) are observational. They describe what was reported, not what would be observed under controlled conditions. Effect sizes should not be interpreted as causal estimates of the "value of reasoning."

Benchmark-specific, not task-general. Our taxonomy (high-value, moderate, negative) applies to the 14 benchmarks studied. Extrapolation to real-world deployment tasks (customer support, document summarization, creative writing) is unsupported by our data and should not be inferred.

Snapshot-specific, not temporally stable. Both model capabilities and API pricing change rapidly. Conclusions about cost–accuracy tradeoffs reflect February 2026 conditions and may not hold six months later.

Aggregation-level, not instance-level. We report benchmark-level averages. Individual query-level variance is unknown from published data and may be substantial, particularly for small-N benchmarks.

Table 11: Practitioner guidance based on observed patterns. All Δ_r values are observational ranges, not causal estimates.

Task Profile	Suggestion	Observed Δ_r	Cost–Benefit
Multi-step math/logic	Reasoning, high effort	+30–81 pp	**Likely favorable**
PhD-level science	Reasoning, med–high	+12–32 pp	**Likely favorable**
Complex SWE tasks	Reasoning, medium	+7–21 pp	**Context-dependent**
General knowledge	Standard model	+2–4 pp	**Unlikely justified**
Factual QA	Standard model	≤ 0 pp	**Likely unfavorable**

15 Causal Limitations

The central limitation of this work is the inability to make causal claims. We identify four primary confounds:

Training confound. Reasoning models receive different training (typically RL-based optimization on reasoning traces) than their standard counterparts. Performance differences may reflect training recipe improvements rather than inference-time reasoning.

Architecture confound. For most provider families (OpenAI, Google), reasoning and standard models are different systems with potentially different architectures, parameter counts, and pre-training data. Only Claude 3.7 (same weights, different inference mode) and DeepSeek V3/R1 (same architecture, different training) provide partial control.

Evaluation confound. Reasoning models are evaluated under different protocols (higher temperature, more samples, tool access) than standard models. These protocol differences alone can account for 5–15 pp of observed accuracy variation.

Selection confound. Providers selectively report results on benchmarks where their reasoning models perform well. The 44.5% missing-data rate is non-random, and the observable data likely overestimates reasoning benefits relative to what a comprehensive evaluation would show.

A fully causal analysis would require experiments that this meta-analysis cannot conduct: same-model, same-prompt, same-sampling-strategy comparisons with reasoning enabled vs. disabled, across multiple models and benchmarks, with raw per-item response data.

16 Future Controlled Experiments Needed

To resolve the ambiguities in this meta-analysis, we identify the following high-priority experiments:

1. **Multi-model token-budget sweep.** Repeat the Claude 3.7 thinking-budget experiment (varying T from 0 to 128K) on at least three additional models (e.g., o3-mini, DeepSeek-R1, Gemini 2.5) and at least five benchmarks, with raw per-item accuracy data reported. This would enable formal model comparison (AIC/BIC) across functional forms.
2. **Protocol-controlled cross-model evaluation.** Evaluate all models under identical prompts, temperature ($T = 0$), single-pass (no majority vote), and no tool access. This would isolate reasoning-specific contributions from evaluation-protocol effects.

3. **Item-level difficulty stratification.** Publish per-item correctness data (not just aggregate accuracy) to enable difficulty-stratified analysis, item-response theory modeling, and proper confidence interval computation.
4. **Contamination auditing.** Systematically test for benchmark contamination by evaluating on held-out variants (e.g., AIME 2026, novel GPQA items) to determine whether reasoning model advantages persist on provably unseen problems.
5. **Latency and throughput analysis.** Report wall-clock latency and tokens-per-second alongside accuracy, enabling cost analysis that accounts for time-to-completion and batch processing efficiency.

17 Limitations

1. **Non-controlled aggregation.** This is a meta-analysis of observational data, not a controlled experiment. Cross-model comparisons confound reasoning mechanisms with training data, architecture, and alignment differences. Only Claude 3.7 provides a same-weights comparison; all other "within-family" comparisons involve different model versions.
2. **Provider reporting bias.** Model providers disproportionately report results on benchmarks where their models excel. Our 44.5% missing-data rate is non-random, biasing the observable data toward favorable results.
3. **Publication bias in benchmarks.** The benchmarks themselves are not a random sample of tasks. They over-represent domains (math, code) where LLM progress is rapid and under-represent domains (open-ended generation, real-world deployment tasks) where evaluation is harder.
4. **Protocol heterogeneity.** Evaluation protocols differ substantially: Claude 3.7 AIME uses parallel sampling + scoring model; Gemini 2.5 Pro reports pass@1; o3/o4-mini distinguish tool vs. no-tool settings. These differences can account for 5–15 pp variation and are not always documented in detail.
5. **Tool/no-tool confounds.** Some reasoning model evaluations (especially on code benchmarks) allow tool use (code execution, web browsing) not available to standard models, inflating observed Δ_r.
6. **Sampling strategy differences.** pass@1, consensus@64, and majority-vote@k produce different accuracy estimates for the same underlying model. We note these where documented but cannot fully normalize.
7. **Small N and missing confidence intervals.** AIME ($n = 30$) has $\sim \pm$ 17 pp 95% CI width at typical accuracy levels. Few official results report confidence intervals, making statistical comparison of model pairs unreliable for benchmarks with $n < 200$.
8. **Hidden reasoning tokens.** OpenAI's o-series thinking tokens are invisible. Our token estimates carry $\pm$20–50% uncertainty, propagating directly into cost analysis.
9. **Latent training data overlap.** Benchmark contamination is documented for GSM8K and MMLU (Scale AI, 2024). AIME and GPQA are newer but not provably clean; reasoning models may benefit from having seen similar problems during training.
10. **Pricing volatility.** API prices are a snapshot (February 2026) and change frequently. Historical trends suggest 2–5× price decreases within 12 months of model launch.
11. **Scaling fit limitations.** Our logarithmic fit is based on $n = 7$ reconstructed data points from one model on one benchmark. The $R^2 = 0.97$ is partly mechanical (two-parameter fit to smooth data) and the fit cannot be statistically distinguished from several alternative functional forms.
12. **Single-author scope.** This analysis was conducted by a single researcher without access to model internals, unpublished data, or institutional review infrastructure. Errors in data extraction are possible.

18 Conclusion

This paper presents a structured meta-analysis of reasoning trace length and reported accuracy across 22 large language models and 14 benchmarks. Our five principal *observational findings*:

Summary of Observations

1. **Large reported accuracy differences on hard tasks.** Reasoning-augmented models outperform standard counterparts by 40–81 pp on competition math and 12–32 pp on PhD-level science, consistently across four provider families. These are observational differences that confound reasoning-specific effects with training improvements.
2. **Diminishing returns in one controlled setting.** For Claude 3.7 on AIME, $A(T) = 13.5 \ln(T + 1) + 18$ ($R^2 = 0.97$, $n = 7$). This fit, while descriptive, cannot be statistically distinguished from alternative functional forms at this sample size. Generalization beyond this setting is not established.
3. **Domain dependence.** Multi-step deduction tasks show large positive differences; factual recall shows negative differences.
4. **Nonlinear estimated cost.** Estimated marginal cost per accuracy point increases substantially near the frontier, though exact magnitudes carry order-of-magnitude uncertainty.
5. **Limited faithfulness.** Published studies report 25–39$\%$ faithful acknowledgment rates, with unfaithful traces being longer.

These patterns, while subject to the substantial limitations documented in Section 17, provide a useful organizing framework for practitioners evaluating reasoning models. The core practical implication is that reasoning-augmented inference appears most beneficial for hard, multi-step deductive tasks and least beneficial—or counterproductive—for factual retrieval and simple classification.

References

Anthropic. Extended thinking for Claude. Technical Report, Anthropic, February 2025.

Anthropic. Reasoning models don't always say what they think. *arXiv:2505.05410*, May 2025.

Apple Machine Learning Research. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *NeurIPS*, 2025.

X. Chen, Y. Wang, Z. Li, et al. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv:2412.21187*, December 2024.

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

Google DeepMind. Gemini 2.5: Our most intelligent AI model. Technical Report, March 2025.

Y. Han et al. SelfBudgeter: Adaptive token allocation for efficient LLM reasoning. *arXiv:2505.11274*, May 2025.

M. Jin et al. The impact of reasoning step length on large language models. *arXiv:2401.04925*, January 2024.

Z. Kang et al. TrimR: Trimming redundant reasoning tokens in LLMs. *arXiv:2505.02152*, May 2025.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

Z. Li et al. Optimal chain-of-thought length follows an inverted-U curve. *arXiv:2503.01141*, March 2025.

Z. Li et al. Let me think! A long chain-of-thought can be worth exponentially many short ones. *arXiv:2505.21825*, May 2025.

N. Muennighoff et al. s1: Simple test-time scaling. *arXiv:2501.19393*, January 2025.

OpenAI. Learning to reason with LLMs. Technical Report, September 2024.

OpenAI. o3 and o4-mini system card. Technical Report, April 2025.

Scale AI. A careful examination of large language model performance on grade school arithmetic. In *NeurIPS*, 2024.

Y. Shen et al. Certainty-guided reasoning in LLMs: A dynamic thinking budget approach. *arXiv:2509.07820*, September 2025.

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*, August 2024.

J. Su et al. Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs. *arXiv:2505.00127*, May 2025.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

X. Wang, J. Wei, D. Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

J. Wei, X. Wang, D. Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

A Extended Results Matrix

Table 12: Complete model–benchmark matrix. All values in %, except Codeforces (Elo). "—" = unreported. 137 of 308 cells (44.5%) are missing.

Model	MATH	AI'24	AI'25	GPQA	MMLU	MMLUPro	HuEv	SWE	CF
GPT-4o	75.9	12	—	53.1	88.7	—	90.2	—	1200
o1	96.4	74.3	79.2	75.7	91.8	—	—	48.9	1891
o1-mini	90.0	70.0	—	—	—	—	—	—	—
o3-mini$_M$	97.3	87.3	86.5	74.9	85.9	—	96.3	49.3	2073
o3$_M$	97.8	91.6	88.9	82.8	92.9	—	87.4	69.1	2706
o4-mini$_M$	97.5	93.4	92.7	77.6	90.0	—	97.3	68.1	2719
Cl. 3.5 Son.	78.3	16	—	65.0	90.5	78.0	93.7	49.0	—
Cl. 3.7$_{64K}$	96.2	80	—	84.8	—	—	—	70.3	—
DS-V3	90.2	39.2	—	59.1	88.5	75.9	82.6	42.0	1134
DS-R1	97.3	79.8	70.0	71.5	90.8	84.0	—	49.2	2029
DS-R1-0528	—	91.4	87.5	—	—	—	—	—	—
Gem. 2.0 Fl.	89.7	35.5	—	—	—	—	—	—	—
Gem. 2.5 Pro	—	92.0	86.7	84.0	89.8	—	—	63.8	—
Llama 405B	73.8	23.3	—	49.0	88.6	73.3	89.0	23.8	759
QwQ-32B	95.6	79.5	65.8	—	—	—	—	—	—
Phi-4-r-plus	—	81.3	78.0	68.9	—	76.0	—	—	—

B Reproducibility Statement

1. **Data availability.** All benchmark scores are traceable to specific published sources cited in the bibliography. The complete data matrix is in Table 12. No proprietary or unpublished data are used.
2. **Extraction methodology.** Data were extracted manually by a single researcher from the following source types: (a) numerical tables in technical reports and system cards (direct transcription), (b) figures in publications (manual digitization from charts, estimated $\pm$2–3 pp uncertainty), (c) API documentation (pricing, token limits). No automated scraping or crowd-sourced extraction was used. A second extraction pass was performed to check for transcription errors; no systematic verification by an independent researcher was conducted.
3. **Table construction.** Tables 3–5 were constructed by matching model–benchmark pairs across sources. Where multiple sources reported the same score, we used the official provider report. Where scores differed across sources (observed in $< 5\%$ of cases), we used the provider-reported value and noted discrepancies in our working notes.
4. **Missing data handling.** Missing cells ("—") indicate that no published result was found. We did not impute, interpolate, or estimate missing values.
5. **Code availability.** No novel code was developed. The analysis consists of manual data extraction, OLS curve fitting, and Pearson correlation—all reproducible with standard statistical software (Python `scipy.optimize.curve_fit`, `numpy.corrcoef`).
6. **Price snapshot.** API prices reflect published rates as of February 15, 2026, verified against official pricing pages. Prices are known to change frequently; readers should verify current rates.
7. **Model versions.** Documented in Table 1. For models with version strings, we use the latest version as of the cited report.
8. **Token accounting.** Visible-thinking models: published counts. Hidden-thinking models (OpenAI o-series): third-party billing analysis, stated uncertainty $\pm$20–50%.

9. **Curve fitting.** OLS on seven data points reconstructed from Figure 3 in Anthropic (2025a). Digitization uncertainty: ±2–3 pp per point.
10. **Assumptions.** (a) Published scores are accurately reported. (b) Token counts are approximately correct. (c) Benchmark items are approximately independent for CI computation. (d) February 2026 prices are representative.
11. **Declaration of no proprietary access.** The author has no access to model weights, internal evaluation logs, unpublished benchmark results, or provider-internal token accounting data for any model analyzed in this paper.

C Notation Reference

Symbol	Definition
$A_{\mathcal{M},\mathcal{B}}$	Reported accuracy of model $\mathcal{M}$ on benchmark $\mathcal{B}$
$\bar{T}_{\mathcal{M},\mathcal{B}}$	Estimated mean total output tokens
η	Token efficiency (accuracy per 1K output tokens)
Δ_r	Observed accuracy difference (reasoning − standard)
$\hat{\Delta}_r$	Token-normalized accuracy difference
$\mathcal{C}$	Cost–accuracy ratio
α, β	Log-fit parameters (fitted, not general)
T^*	Optimal token budget (model-dependent)
λ	Cost sensitivity parameter
pp	Percentage points

D Revised Title Options

1. *Reasoning Trace Length and Accuracy in Large Language Models: A Structured Meta-Analysis of Published Benchmarks* (primary; accurately reflects methodology)
2. *When Does Thinking Help? A Cross-Family Meta-Analysis of Reasoning Model Performance, Cost, and Faithfulness*
3. *The Accuracy–Cost Frontier of Inference-Time Reasoning: Patterns from 22 Models and 14 Benchmarks*

E Reviewer Rebuttal Strategy

Anticipated criticism: "The scaling analysis is a curve fit to 7 digitized points. This is not a finding." We agree this is the weakest empirical component. We have (a) explicitly labeled it as a descriptive fit, not a law; (b) added residual analysis; (c) discussed alternative functional forms (power, sqrt, sigmoid) that would fit equally well; (d) stated falsifiability conditions; and (e) restricted all derived quantities (marginal gain table) to this specific model–benchmark pair. We retain the fit because it formalizes the qualitatively obvious diminishing-returns pattern in the only controlled token-budget data publicly available.

Anticipated criticism: "Cross-model comparisons are not commensurable." Acknowledged in Section 3 (metric normalization), Section 10 (confounding variables), Section 15 (evaluation confound), and Section 17 (protocol heterogeneity, item 4). We estimate 5–15 pp uncontrolled protocol variation and flag non-pass@1 evaluations with † throughout. We explicitly do not claim commensurability.

Anticipated criticism: "The contribution is a literature survey, not a research paper." We position this as a structured meta-analysis, not as a novel empirical study. Our contribution is in systematization: formal metric definitions, cross-family aggregation with explicit missing-data documentation, sensitivity analysis, and a deployment framework. We believe this fills a gap—no prior work has aggregated reasoning model performance across all five major provider families with formal efficiency metrics and cost analysis—while being transparent about the limitations of observational aggregation.

Anticipated criticism: "Statistical claims on n=4 and n=5 are not credible." We have downgraded all small-n correlations to "purely descriptive" with explicit statements that "no statistical inference should be drawn." These values are reported as compact summaries of observed monotonic trends, not as evidence for underlying relationships.

Anticipated criticism: "44% missing data could reverse findings." Addressed in the new missing-data sensitivity subsection (Section 10). We discuss worst-case scenarios and acknowledge that the direction of missing-data bias cannot be determined.