COST-OF-PASS: AN ECONOMIC FRAMEWORK FOR EVALUATING LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The widespread adoption of AI systems in the economy hinges on their ability to generate economic value that outweighs their inference costs. Evaluating this tradeoff requires metrics that account for both performance and costs. We propose a framework grounded in production theory for evaluating language models by combining accuracy and inference cost. We introduce *cost-of-pass*, the expected monetary cost of generating a correct solution. We then define the *frontier cost-of-pass* as the minimum cost-of-pass achievable across available models *or the human-expert*, using the approximate cost of hiring an expert. Our analysis reveals distinct economic insights. *First*, lightweight models are most cost-effective for basic quantitative tasks, large models for knowledge-intensive ones, and reasoning models for complex quantitative problems, despite higher per-token costs. *Second*, tracking this frontier cost-of-pass over the past year reveals significant progress, particularly for complex quantitative tasks where the cost has roughly halved every few months. *Third*, to trace key innovations driving this progress, we examine counterfactual frontiers—estimates of cost-efficiency without specific model classes. We find that innovations in lightweight, large, and reasoning models have been essential for pushing the frontier in basic quantitative, knowledge-intensive, and complex quantitative tasks, respectively. *Finally*, we assess the cost-reductions from common inference-time techniques (majority voting and self-refinement), and a budget-aware technique (TALE-EP). We find that performance-oriented methods with marginal performance gains rarely justify the costs, while TALE-EP shows some promise. Overall, our findings underscore that complementary model-level innovations are the primary drivers of cost-efficiency, and our economic framework provides a principled tool for measuring this progress and guiding deployment.

1 INTRODUCTION

The recent progress in generative AI, particularly language models (LMs), has sparked significant interest in their potential to transform industries, automate cognitive tasks, and reshape economic productivity (Brynjolfsson et al., 2025; Eloundou et al., 2024; Acemoglu, 2024). The widespread adoption of these AI systems in the economy hinges on whether the economic benefits generated by the tasks they can perform outweigh the associated inference costs, and whether those inference costs are lower than the cost of equivalent human labor. Consequently, two priorities have emerged at the forefront of LM research: advancing capabilities and reducing costs. These goals, however, often involve trade-offs with more powerful models or test-time techniques that offer higher accuracy at the expense of greater computational and monetary cost (Chen et al., 2024; Parashar et al., 2025; Madaan et al., 2023; Wang et al., 2023; Kapoor et al., 2024). While standard metrics capture accuracy or other system capabilities, they fail to account for cost, leading to an incomplete picture of progress. Ultimately, what matters to the users is not just raw capability, but the value delivered relative to cost and the standard has been to interpret and report these separately. As the ecosystem of models grows, it is essential to assess new models not in isolation, but in the context of a broader ecosystem, where marginal improvements may or may not justify higher costs, and do so in an easy-to-interpret manner.

To systematically investigate the trade-off between cost and performance and analyze the LM ecosystem as a whole, we draw insights from a well-established and foundational framework from economics: production frontiers. Economists have long studied these frontiers, which map a set of inputs to *the maximum output attainable under a given technology* (Farrell, 1957). In Farrell's original

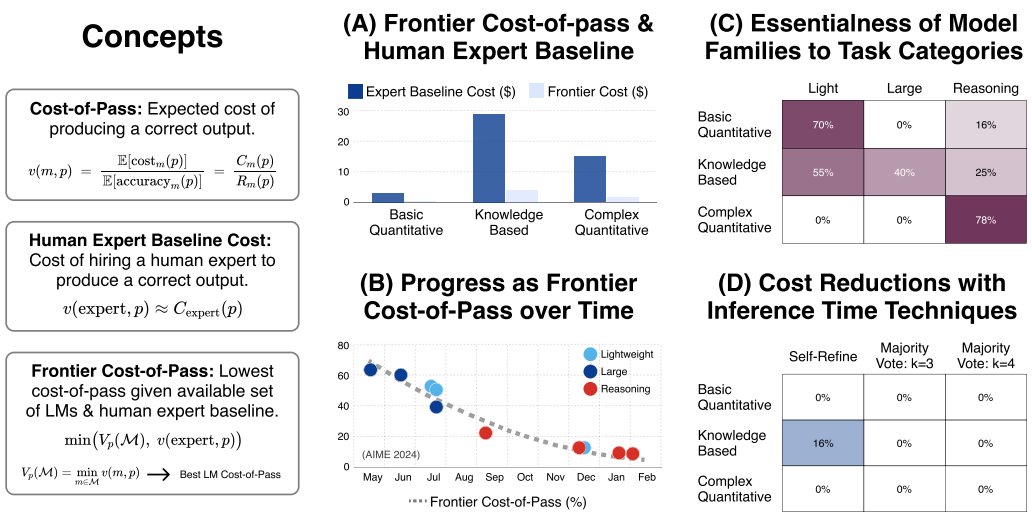

Figure 1: Highlights of the cost-of-pass framework and empirical analyses. Core concepts (left) set foundations for: **(A)** Comparing the Human Expert Baseline to the frontier achieved by the single most effective LM per task category. **(B)** Tracking the reduction in frontier cost-of-pass over time, indicating progress driven by new model releases (color-coded by family). **(C)** Quantifying the essential contribution of each model family: lightweight (less than $1 per million tokens), large, and reasoning; to the current cost-efficiency frontier, measured by the percentage of each family's contribution. **(D)** Assessing the economic benefit (relative cost reduction) achieved by applying common inference-time techniques over the baseline model frontier (which rarely results in meaningful gains).

formulation, a producer is *technically efficient* if no input can be reduced without lowering output, and *price efficient* if the input mix minimizes cost given input prices. Together, these conditions yield the lowest possible cost per unit of output. Extending this framework, Aigner et al. (1977) introduced stochastic frontier production functions, in which the relationship between inputs and output is modeled as stochastic rather than deterministic, practically accounting for potential defective outputs that do not pass evaluation criteria due to factors beyond the producer's control.

These economic concepts are highly relevant to modern LMs, which inherently function as stochastic producers: for a given input, they yield a desired output (e.g., a correct solution) stochastically (Brown et al., 2024). Common practices such as employing scaffolds or more computationally intensive inference techniques (Snell et al., 2024; Madaan et al., 2023; Wang et al., 2023) represent efforts to manipulate this production process. These strategies seek to increase the probability of success but typically do so at the expense of higher computational cost, directly mirroring the economic trade-offs inherent in production efficiency. Motivated by these parallels and the economic goal of minimizing cost per successful output under uncertainty, we develop a quantitative framework tailored to LMs.

We summarize our contributions as follows.

Concepts. We introduce *cost-of-pass* (§2.2), which quantifies the expected monetary cost to achieve a successful output for a given problem. Building on this concept and incorporating a human-expert cost baseline, we define the *frontier cost-of-pass* (§2.4) as the minimum achievable cost-of-pass across all available options (LMs and human-expert) for that problem. We show these reveal distinct economic niches for model families (e.g., lightweight *vs.* reasoning models) on different tasks, which accuracy comparisons alone obscure (§3.2).

Tracking progress with frontier cost-of-pass. Using the cost-of-pass and frontier cost-of-pass, we analyze economic improvements across three task categories from May 2024 to February 2025. We observe an exponential decrease in frontier cost-of-pass across all tasks, though the trends vary. Notably, we observe that, over the past year, the expected cost of generating a correct solution to complex quantitative problems has been cut in half every few months. We find that the frontier cost-of-pass is driven primarily by lightweight models and reasoning models (§3.3).

Counterfactual frontier in the absence of model families. We show that our analysis reveals the complementary roles of different model types in driving recent progress. Innovations in lightweight

models have been instrumental in reducing costs on basic quantitative tasks. Large models, by contrast, have been most impactful for knowledge-based benchmarks like `GPQA-Diamond` (Rein et al., 2024). Meanwhile, reasoning models have been central to advances on complex quantitative reasoning challenges such as `AIME` (MAA, 2024) and `MATH` (Hendrycks et al., 2021) (§ 3.4).

Impact of post-hoc inference time techniques. We observe that common test-time techniques such as self-refinement (Madaan et al., 2023) and majority voting (self-consistency; Wang et al., 2022) to improve performance offer either limited or no economic benefits, while a budget-aware technique TALE-EP (Han et al., 2024) delivers some benefits. These indicate that the recent reductions in frontier cost-of-pass have been mostly driven by model-level innovations (§ 3.5).

2 SETUP

2.1 ECONOMIC THEORY OF PRODUCTION EFFICIENCY

Classical production theory examines how producers efficiently convert inputs (resources) into outputs. A central concern is understanding the maximum output attainable with a given set of inputs, or conversely, the minimum inputs (and thus cost) required to achieve a specific target output level.

Consider a set of producers $\mathcal{F} = \{f_i\}_{i=1}^{n}$ such that each producer $f_i \in \mathcal{F}$ can transform an input vector $\mathbf{x} \in \mathbb{R}^k_{\geq 0}$ (e.g., quantities of different resources) into an output. The inputs used by producer f_i have associated prices, represented by a price vector $\mathbf{w_i} \in \mathbb{R}^k_{\geq 0}$. When focusing on achieving a specific target output level, say u units, economists are interested in the *frontier cost* V_u. This represents the absolute minimum monetary cost required to produce at least u units, considering all input vectors $\mathbf{x}$ capable of achieving this output across all available producers with their respective pricings. This frontier cost is formally defined as:

$$V_u = \min_{f_i \in \mathcal{F}} \left\{ \mathbf{w}_i^\top \mathbf{x} \,\middle|\, f_i(\mathbf{x}) \geq u \right\}, \tag{1}$$

Farrell (1957) used these core concepts to formalize definitions for technical and price efficiency in a production ecosystem for producers. Critically, Aigner et al. (1977) extended this framework to handle *stochastic* production functions, where output is probabilistic for a given input.

Building on this economic foundation, we adapt the core concept of a frontier cost (V_u) to represent the minimum achievable cost for obtaining a correct solution using LMs. To better reflect LM behavior, which is inherently stochastic, we incorporate this variability into our cost-efficiency metric. This aligns our framework with core production concepts and enables assessment of the economic impact of stochastic LM producers.

2.2 COST-OF-PASS: AN EFFICIENCY METRIC FOR LMS

Here we instantiate the economic framework for language models (LMs). Consider a specific *problem* p, where the unit of production is a correct solution. We define a *model* m as an inference pipeline using an LM, acting as a stochastic producer. Two quantities characterize its efficiency on problem p:

$$R_m(p) = \text{Prob. of } m \text{ producing a correct answer on } p,$$
$$C_m(p) = \text{Expected cost of one inference attempt by } m \text{ on } p.$$

In the context of LMs, the inputs $\mathbf{x}$ correspond to resources like prompt and generated tokens, while the input prices $\mathbf{w}$ represent the costs per token charged by the provider. The total cost of these inputs for a single inference attempt by model m on problem p is captured by $C_m(p)$, effectively instantiating the term $\mathbf{w}^\top \mathbf{x}$ from the theory in the previous section.

Since the model output is stochastic, the expected number of attempts to obtain the first correct solution is $1/R_m(p)$, assuming independent trials. This yields the ***cost-of-pass***, defined as the expected monetary cost to obtain one correct solution for problem p:

$$v(m, p) \;=\; \frac{C_m(p)}{R_m(p)}. \tag{2}$$

The cost-of-pass integrates both performance ($R_m(p)$) and cost ($C_m(p)$) into a single economically interpretable metric: it quantifies how efficiently financial resources are converted into correct outputs.

This formulation mirrors classical production theory, where the goal is to assess the cost of achieving a specific target output (Farrell, 1957); in our case, the target is a correct solution. When a model cannot produce one ($R_m(p) = 0$), the cost-of-pass becomes infinite, appropriately signaling infeasibility.

2.3 The LM Frontier Cost-of-Pass

While cost-of-pass (§ 2.2) evaluates a single model's efficiency, understanding the overall state of LM capabilities for a given problem requires assessing the collective performance of the entire available LM ecosystem. Therefore, analogous to the frontier cost V_u (Eq. 1), we define the ***LM frontier cost-of-pass*** for problem p as the minimum cost-of-pass achievable using any available LM strategy m from the set $\mathcal{M}$:

$$V_p(\mathcal{M}) = \min_{m \in \mathcal{M}} v(m, p). \tag{3}$$

$V_p(\mathcal{M})$ quantifies the minimum expected cost to solve problem p using the most cost-effective model currently available within the set $\mathcal{M}$. If no LM in $\mathcal{M}$ can solve p (i.e., $R_m(p) = 0$ for all $m \in \mathcal{M}$), then $V_p(\mathcal{M}) = \infty$.

2.4 Grounding Evaluation: Estimated Human-Expert Baseline

The LM frontier cost-of-pass $V_p(\mathcal{M})$ reveals the best LM performance but lacks context: it does not show if LMs are economically advantageous over human labor. Moreover, the LM frontier cost-of-pass can be infinite if no LM succeeds. To address both, we introduce *human-expert baseline* as a reference point, by considering a human-expert annotator as a specific strategy: m_{expert}. Let $\mathcal{M}_0 = \{m_{\text{expert}}\}$ represent this baseline set. We assume experts typically achieve near-perfect correctness ($R_{\text{expert}}(p) \approx 1$) for tasks they are qualified for. Thus, the cost-of-pass for a qualified expert is approximately their labor cost per problem:

$$v(\text{expert}, p) \approx C_{\text{expert}}(p). \tag{4}$$

The estimation of $C_{\text{expert}}(p)$ involves considering required expertise, time per problem, and appropriate compensation rates (detailed in § 2.6.1). By incorporating this baseline, we define the *frontier cost-of-pass* for problem p, considering both LMs ($\mathcal{M}$) and the human-expert alternative ($\mathcal{M}_0$):

$$V_p(\mathcal{M} \cup \mathcal{M}_0) = \min\big(V_p(\mathcal{M}),\ v(\text{expert}, p)\big). \tag{5}$$

This frontier cost-of-pass represents the true minimum expected cost to obtain a correct solution for problem p using the best available option, whether it's an LM or a human. Crucially, $V_p(\mathcal{M} \cup \mathcal{M}_0)$ is always finite (assuming finite human-expert cost and capability).

2.5 Measuring Progress and Value Gain

To track improvements against the best available option over time, let $\mathcal{M}_t$ denote the *total set of available strategies* at time t, encompassing both the set of LM strategies released up to time t *and* the human-expert baseline $\mathcal{M}_0$, that is, $\mathcal{M}_t = \{m_{\leq t}\} \cup \mathcal{M}_0$. The frontier cost-of-pass achievable at time t can be calculated as:

$$V_p(\mathcal{M}_t) = \min_{m \in \mathcal{M}_t} v(m, p). \tag{6}$$

As new LM models $\{m_t\}$ are released, the set expands such that $\mathcal{M}_t = \mathcal{M}_{t-1} \cup \{m_t\}$. Consequently, the frontier cost-of-pass $V_p(\mathcal{M}_t)$ forms a *non-increasing* sequence over time t, tracking the reduction in the minimum cost needed to solve a particular problem p.

To quantify the economic impact of new developments, we define the *gain*. When a new set of models $\{m_t\}$ becomes available at time t (often a single model), the gain for problem p is the reduction it causes in the frontier cost-of-pass:

$$G_p(\{m_t\}, \mathcal{M}_{t-1}) = V_p(\mathcal{M}_{t-1}) - V_p(\mathcal{M}_{t-1} \cup \{m_t\}). \tag{7}$$

Note that G_p measures how much cheaper the new model(s), $\{m_t\}$, make solving p compared to prior best options, including humans. Hence, a large G_p value indicates a significant economic contribution in solving p. This notion underlies our experiments, analyzing the value generated by models relative to the human baseline and tracking the evolution of the overall frontier.

Extending to a distribution. Although measuring frontier cost-of-pass and value gain for individual problems can be informative, particularly through a fine-grained perspective, we often care about more than a single instance. Let $P = \{p_i\}_{i=1}^n$ be n problems drawn i.i.d. from D. We treat P as the empirical distribution that puts mass $1/n$ on each element. We can then extend our definitions for such a distribution through the following:

$$V_{p\sim D}(\mathcal{M}_t) \approx \mathbb{E}_{p\sim P}[V_p(\mathcal{M}_t)], \tag{8}$$

$$G_{p\sim D}(\{m_t\}, \mathcal{M}_{t-1}) \approx \mathbb{E}_{p\sim P}[G_p(\{m_t\}, \mathcal{M}_{t-1})]. \tag{9}$$

2.6 Estimating the Economic Efficiency

To operationalize our overall framework for any given distribution of problems, we introduce the following recipe:

(1) **Estimate success rates.** For each model-problem pair (m, p), generate a number of independent attempts to approximate $R_m(p)$. Use the same prompt and model settings across these attempts, varying only factors necessary to ensure independence (e.g., internal sampling randomness).

(2) **Estimate per-attempt cost.** Track the average number of tokens (prompt + generation) consumed per attempt, multiply by the current token price (which can differ by model provider or usage level), and add any extra charges (e.g., third-party API calls, external reasoning modules, etc.). This sum yields $C_m(p)$.

(3) **Compute cost-of-pass.** For each model m, calculate $v(m, p) = C_m(p)/R_m(p)$. ($R_m(p) = 0$ yields $v(m, p) = \infty$.)

(4) **Determine frontier cost-of-pass.** Estimate human-expert cost $v(\text{expert}, p)$ (see below). Find $V_p(\mathcal{M} \cup \mathcal{M}_0)$ for a given set of strategies $\mathcal{M}$.

(5) **Analyze over benchmarks.** Aggregate $V_p(\mathcal{M})$ across problems $p \sim D$ to estimate $V_{p\sim D}(\mathcal{M}_t)$. Track progress over time (for $\mathcal{M}_t$) and compute gain $G_{p\sim D}$ for new models.

2.6.1 Estimating Human-Expert Cost

To estimate $v(\text{expert}, p)$, the plausible cost of obtaining a correct human-expert answer, we systematically determine the required qualifications, appropriate hourly compensation, and average time for a typical problem p per dataset. We determine these quantities based on a hierarchy of evidence by prioritizing the dataset's creation process or associated studies (e.g., reported annotation pay/time (Parrish et al., 2022)). When direct data is absent, we leverage findings from closely related work (Zhang et al., 2024) or infer parameters from the dataset's context (e.g., deriving time-per-problem from contest rules (Art of Problem Solving, 2023)). Compensation rates are informed by reported study payments (Rein, 2024) or relevant market rates for comparable expertise (e.g., specialized tutoring rates (TutorCruncher, 2025; Wyzant Tutoring, 2025)).*

3 Experiments

3.1 Experiment Setup

Models. We consider three categories of models:

(1) ***Lightweight*** **models:** We use the per-token cost as a proxy and select models with a cost less than $1 per million input and output tokens (see Table 4): Llama-3.1-8B (Grattafiori et al., 2024), GPT-4o mini (OpenAI, 2024), and Llama-3.3-70B (Meta-AI, 2024).

(2) ***Large*** **models:** We select large general-purpose LMs: Llama-3.1-405B (Grattafiori et al., 2024), Claude Sonnet-3.5 (Anthropic, 2024), and GPT-4o (Hurst et al., 2024).

(3) ***Reasoning*** **models:** We select models with special reasoning post-training, including OpenAI's o1-mini (OpenAI et al., 2024), o1 (OpenAI et al., 2024), and o3-mini (OpenAI, 2025), as well as DeepSeek R1 (Guo et al., 2025).

*The full derivation, justification, and sources for our approach are detailed in Appendix A. The resulting estimates are in Table 3.

Model Category	Basic Quantitative		Knowledge Based		Complex Quantitative	
	2-Digit Add.	GSM8K	BBQ	GPQA Dia.	MATH 500	AIME24
Lightweight Models						
Llama-3.1-8B	4.8e−5	0.19	2.7e−2	18.58	3.38	15.33
GPT-4o mini	5.4e−5	0.22	1.3e−2	25.38	2.06	14.67
Llama-3.3-70B	1.6e−4	0.16	7.4e−3	18.58	1.31	10.67
Large Models						
Llama-3.1-405B	6.9e−4	0.14	6.7e−3	10.43	1.13	8.67
Claude Sonnet-3.5	2.1e−3	0.19	6.4e−3	14.06	2.54	14.67
GPT-4o	2.3e−3	0.17	6.2e−3	14.07	0.96	14.01
Reasoning Models						
OpenAI o1-mini	5.4e−3	0.17	1.3e−2	12.27	0.50	4.80
OpenAI o1	1.9e−2	0.22	4.3e−2	8.07	0.90	2.85
DeepSeek-R1	1.8e−3	0.17	1.5e−2	14.57	0.21	3.41
OpenAI o3-mini	1.1e−3	0.11	1.1e−2	8.18	0.76	2.03

Table 1: Frontier dollar cost-of-pass per model / dataset. Each entry is the expected dollar cost of a problem $p \sim D$ with the presence of the model m and a human expert: $V_{p\sim D}(\{m\} \cup \mathcal{M}_0)$. Per column, the 3 entries with the lowest value (i.e. best frontier cost-of-pass) have blue highlights. Different model families emerge as cost-effective at different task categories, highlighting the strengths of our evaluation.

Within each category, we select three to four representative models released between the second half of 2024 and early 2025. To preserve the integrity of our temporal analysis, we prioritize the earliest stable releases and exclude research previews or experimental versions.

Datasets. We evaluate models across three sets of tasks:

(1) ***Basic quantitative* tasks:** These involve basic numerical reasoning. We include an arithmetic dataset (`Two Digit Addition`) to assess basic numerical computation, and `GSM8K` (Cobbe et al., 2021) to evaluate multi-step grade-school level problem solving.

(2) ***Knowledge-based* tasks:** These require recalling and reasoning over factual knowledge. We include a scientific knowledge-intensive question answering task (`GPQA-Diamond` (Rein et al., 2024)) to evaluate models' ability to recall and utilize complex scientific facts, and a bias benchmark (`BBQ` (Parrish et al., 2022)) to evaluate whether models rely on stereotypical knowledge or can disambiguate factual responses from biased defaults.

(3) ***Complex quantitative reasoning* tasks:** These require complex mathematical reasoning and problem solving. We use `MATH-500` (Hendrycks et al., 2021; Lightman et al., 2023) to assess models on competition-level maths problems, and `AIME-24` (MAA, 2024) to evaluate performance on challenging problems from the 2024 American Invitational Mathematics Examination.

Evaluation protocol. All implementation details including model API providers, per-token pricing, prompt template, sampling budget, and accuracy/cost calculation details are shared in Appendix B.

3.2 Frontier Cost-of-Pass with a Single Model

In this experiment, we aim to quantify the economic value each model m generates on different distributions of problems $p \sim D$. For this, we take human-expert as a baseline and quantify the frontier cost-of-pass of a problem in the presence of the model m: $V_{p\sim D}(\{m\} \cup \mathcal{M}_0)$.

The results in Table 1, highlighting the lowest three instances per dataset, show that our frontier cost-of-pass effectively captures how different model families offer economic advantages across various task categories. We find that lightweight models yield the lowest frontier cost-of-pass on basic quantitative tasks, such as `Two Digit Addition`. This outcome aligns with the observation that all model families achieve high accuracy on this dataset (see Table 5), which in turn makes the least expensive models appear most cost-effective. In contrast, for knowledge-based tasks, larger models achieve a lower frontier cost-of-pass compared to lightweight ones. While the reasoning models, such as o1, are priced significantly more expensively compared to both large and lightweight models, they lead to significant performance improvements, which, overall, result in reductions in the cost-of-pass mainly in complex quantitative tasks.

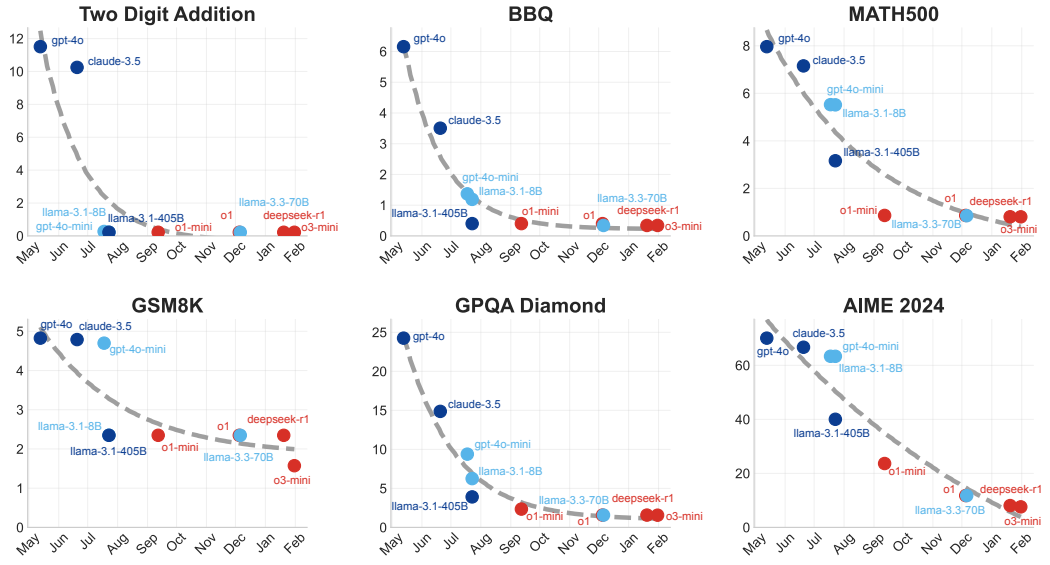

Figure 2: The frontier dollar cost-of-pass (i.e. $V_{p\sim D}(\mathcal{M}_t)$ steadily decreases with new model releases, spanning models released between May 2024 and February 2025. Y-axes are normalized (divided by $V_{p\sim D}(\mathcal{M}_0)$, shown in percentage (%)).

In contrast, when either task performance ($R_m(p \sim D)$) or cost ($C_m(p \sim D)$ is solely taken into account (Tables 5 and 6) such metrics tend to favor either reasoning or lightweight models respectively due to their significant edge per criteria, without assessing the nuances in the economic impact they induce. This effectively highlights the sophistication of our metric and evaluation framework.

3.3 Tracking Frontier Cost-of-Pass with New Releases

In this experiment, we track the improvements on the frontier cost-of-pass for a problem. Figure 2 shows the trends of the cumulative gain per dataset ($V_{p\sim D}(\mathcal{M}_t)$), each updated by the corresponding model release ($\mathcal{M}_{t-1} \cup \{m_t\}$). We observe a steady decline in the frontier cost-of-pass for complex quantitative tasks. In contrast, knowledge-based and basic quantitative tasks typically exhibit a sharp initial drop in frontier cost-of-pass with the early releases of models, followed by a plateau. To quantify the cost reduction trends, we empirically fit an exponential decay curve of the form:

$$V_p(M_t) \approx a\, e^{-b\, t} + c, \tag{10}$$

where t denotes time in months since the first model release, and a, b, and c are fit parameters. From this, we compute the time for the exponential component of the cost to drop by 50%: $T_{1/2} = \ln(2)/b$. Using this formulation, we find that for complex quantitative tasks, between May 2024 and February 2025, the frontier cost-of-pass for `MATH-500` halved approximately every 2.6 months, whereas for `AIME-2024`, the halving time was 7.1 months; indicating consistent cost reductions over the past year.

3.4 Essentialness of Model Families: Counterfactual Frontier Cost-of-Pass

Section 3.3 showed the frontier cost-of-pass decreasing over time with new model releases. To understand which model families were most critical to this progress, we conduct a counterfactual analysis that quantifies the impact of removing each family. Defining $\mathcal{M}_g$ as a family of models (lightweight, large, or reasoning), we measure the counterfactual contribution of family g on dataset D by calculating the relative improvement in frontier cost-of-pass attributable to its inclusion:

$$\frac{G_{p\sim D}(\mathcal{M}_g, \mathcal{M}_T \setminus \mathcal{M}_g)}{V_{p\sim D}(\mathcal{M}_T \setminus \mathcal{M}_g)}. \tag{11}$$

Here, $\mathcal{M}_T$ includes all models used in our experiments. This metric represents the relative improvement in the final frontier cost-of-pass $V_{p\sim D}(\mathcal{M}_T)$ attributable to the model family $\mathcal{M}_g$, with higher values indicating greater essentialness of that family for achieving the current frontier.

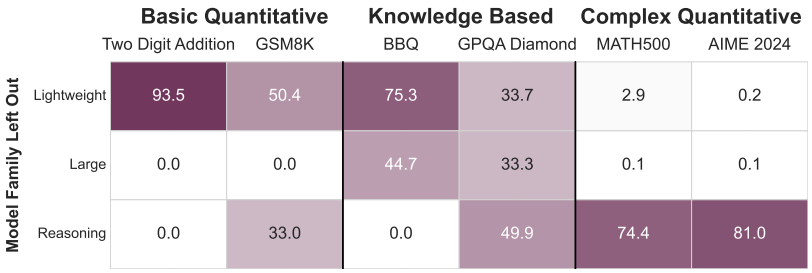

Figure 3: The relative improvement (%) in frontier cost-of-pass attributable to each model family g, calculated under a counterfactual setting where $\mathcal{M}_g$ is removed. Higher values signify greater essentialness for maintaining the current frontier.

Figure 3 illustrates our main findings, revealing distinct roles across model families. Lightweight models help reduce the frontier cost-of-pass on basic quantitative tasks, while large models are only essential in knowledge-intensive tasks. Reasoning models play a key role in advancing the frontier for complex quantitative reasoning and also improve performance on `GPQA-Diamond`, as well as `GSM8K`, which benefits from small reasoning models like `o3-mini`.

These findings highlight that progress on different task types is driven by different model paradigms. While large models have brought clear gains on knowledge-intensive tasks (e.g. `GPQA`), improvements in cost-efficiency, especially in more quantitative domains, appear largely driven by advances in lightweight and reasoning models. Together, these suggest that the current cost-efficiency frontier, as reflected in our framework, is shaped mainly by (i) lightweight models and (ii) reasoning models.

3.5 Impact of Inference Time Techniques on Frontier Cost-of-Pass

We now assess whether common inference-time techniques provide meaningful economic benefits. Specifically, we ask: is it cost-effective to improve model performance through these techniques, compared to relying on the models' baseline performance? To explore this, we focus on the set of lightweight and large models, denoted by $\mathcal{M}_L$. First, we determine the frontier cost-of-pass achieved

Inference Time Technique	Basic Quantitative		Knowledge Based		Complex Quantitative	
	Two Digit Addition	GSM8K	BBQ	GPQA Diamond	MATH500	AIME24
TALE-EP	1.5	66.6	24.5	50	0.2	16.6
Self-Refinement	0	0	6.7	24.9	0	0
Majority Voting (k=3)	0	0	0	0	0	0
Majority Voting (k=4)	0	0	0	0	0	0

Table 2: Relative performance gains (%) from different inference time techniques across datasets.

by $\mathcal{M}_L$ without any modifications. We then apply a given inference-time technique uniformly across all models in $\mathcal{M}_L$, yielding a modified set $\mathcal{M}_L^*$. The gain from this technique, measured relative to the original frontier cost-of-pass, can be computed as follows:

$$\frac{G_{p\sim D}(\mathcal{M}_L^*,\ \mathcal{M}_L)}{V_{p\sim D}(\mathcal{M}_L)}. \tag{12}$$

We consider two popular techniques: self-refinement Madaan et al. (2023) and majority voting (a.k.a. self-consistency; Wang et al., 2023), with 3 and 4 votes. Moreover, we evaluate a budget-aware inference-time technique: TALE-EP Han et al. (2024) as well. As shown in Table 2, self-refinement shows some economic benefit on knowledge-intensive tasks, considerably 24.9% improvement on `GPQA-Diamond`. In contrast, majority voting (despite potentially enhancing accuracy) does not offer relative economic improvement across the tested models and datasets. Meanwhile, the budget-aware technique contributes meaningfully in many more of the tasks to reducing the frontier cost-of-pass.

Collectively, these findings suggest that, for the evaluated techniques, the costs by performance-oriented methods often outweigh accuracy gains when measured by the frontier cost-of-pass. By contrast, TALE-EP (conditioning generation on a self-predicted token budget) yields visible reductions on a subset of tasks, though benefits are uneven. This implies that such common inference-time approaches may currently offer limited economic benefits within our evaluation framework.

4 Related Works

Economic perspectives and broader impacts. The efficiency of LMs carries significant economic implications, as they are viewed as general-purpose technologies impacting productivity and labor (Eloundou et al., 2024; Brynjolfsson et al., 2025). Complementary economic analyses explore provider strategies regarding pricing and product design Bergemann et al. (2025), and user-side decision-making involving ROI, token costs, and success probabilities.

Our cost-of-pass metric serves as a crucial bridge between these technical realities of model performance and their economic consequences. By providing a fundamental measure, the expected monetary cost to successfully complete a task, it allows for quantifying the economic contribution of specific AI systems and informs rational model selection for achieving economic viability, and provides quantitative perspective on the economic evolution of the LM ecosystem.

LM resource consumption, efficiency optimization and benchmarking. Research increasingly recognizes the importance of LM resource consumption and efficiency. Studies have quantified operational costs like tokens (Chen et al., 2023) and energy (Maliakel et al., 2025), revealing task-dependent performance and potential diminishing returns from high expenditure (Miserendino et al., 2025). This focus has intensified with the rise of reasoning methodologies (Sui et al., 2025) and inference-time techniques (e.g., Madaan et al. (2023); Wang et al. (2023)), which often trade increased computational cost for potential accuracy gains.

Concerns like "overthinking," where lengthy processing fails to improve results (Chen et al., 2024; Cuadron et al., 2025), have spurred efforts to optimize resource use through methods like dynamic token budgeting (Han et al., 2025), specialized training (Arora & Zanette, 2025), prompt engineering (Xu et al., 2025; Aytes et al., 2025) or researching optimal reasoning lengths (Wu et al., 2025; Yang et al., 2025). Concurrently, evaluation methodologies have evolved beyond pure accuracy or correctness measures.

Recognizing its insufficiency, researchers have incorporated cost via fixed budgets (Wang et al., 2024), performance heuristics (McDonald et al., 2024), or non-monetary metrics like conciseness (Nayab et al., 2024). Kapoor et al. (2024) strongly advocated for using real dollar costs and accounting for stochasticity—factors central to our approach. Benchmarking efforts have also highlighted diminishing returns from simply scaling inference computation (Parashar et al., 2025). While these works underscore the need for cost-aware analysis, they often rely on specific constraints (e.g., fixed budgets) or heuristic metrics.

Our cost-of-pass framework seeks to advance this by providing a single, interpretable metric grounded in economic production principles, offering a unified way to assess the economic viability of different models and techniques without predefined budget assumptions or proxy metrics.

5 Conclusion

We introduced an economic framework designed to evaluate language models by integrating their performance with inference cost. Drawing from production theory, we conceptualize language models as stochastic producers, and assess their efficiency using our proposed *cost-of-pass* metric, which measures the expected cost per correct solution. Our analysis utilizes this metric alongside the *frontier cost-of-pass*, defined as the minimum achievable cost compared to an human expert baseline. This approach reveals distinct economic roles played by different model classes. For instance, retrospective and counterfactual evaluations demonstrate that lightweight models primarily drive efficiency on basic tasks, whereas reasoning models are essential for complex problem-solving. Critically, our findings show that common inference-time techniques typically increase the *cost-of-pass*, thus failing to provide net economic benefits when compared to the progress made by improving the underlying models themselves. We discuss the limitations of our methodology, outline directions for future work, and consider practical implications of our framework in Appendix D. Taken together, these insights underscore the value of our framework in offering a principled foundation for measuring language model innovation in economic terms. It serves as a valuable tool for guiding model selection and aligning AI development with real-world value.

REFERENCES

1st grade 4th quarter expectations – fast facts timed tests. Elementary School Curriculum Note (online PDF), 2021. URL https://content.myconnectsuite.com/api/documents/c5424d1247714174b26ed5fb00e4ebc3.pdf. States 20–25 addition problems should be solved in 1 minute (2–3 sec each). Accessed 2025-05-14.

Daron Acemoglu. The Simple Macroeconomics of AI. NBER Working Papers 32487, National Bureau of Economic Research, Inc, May 2024. URL https://ideas.repec.org/p/nbr/nberwo/32487.html.

Dennis Aigner, C.A.Knox Lovell, and Peter Schmidt. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37, 1977. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(77)90052-5. URL https://www.sciencedirect.com/science/article/pii/0304407677900525.

Anthropic. Claude 3.5 sonnet announcement, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 13 Feb. 2025.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

Art of Problem Solving. American Invitational Mathematics Examination (AIME) Format. AoPS Wiki (aops.com), 2023. URL https://artofproblemsolving.com/wiki/index.php/American_Invitational_Mathematics_Examination. States AIME is 15 questions in 3 hours (12 min per problem). Accessed 2025-05-14.

Simon A Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. *arXiv preprint arXiv:2503.05179*, 2025.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The economics of large language models: Token allocation, fine-tuning, and optimal pricing. *arXiv preprint arXiv:2502.07736*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, pp. qjae044, 2025.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. Dataset licensed under the MIT License.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024.

Michael James Farrell. The measurement of productive efficiency. *Journal of the royal statistical society: series A (General)*, 120(3):253–281, 1957.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this API serving? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QCDdI7X3f9.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. License: Llama 3 Community License.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. License: MIT License, see https://huggingface.co/deepseek-ai/DeepSeek-R1/blob/main/LICENSE.

Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning, 2025. URL https://arxiv.org/abs/2412.18547.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe. Dataset licensed under the MIT License.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

MAA. American Invitational Mathematics Examination (AIME). https://maa.org/maa-invitational-competitions/, 2024. Accessed: 2025-05-14.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Paul Joe Maliakel, Shashikant Ilager, and Ivona Brandic. Investigating energy efficiency and performance trade-offs in llm inference across tasks and dvfs settings. *arXiv preprint arXiv:2501.08219*, 2025.

Tyler McDonald, Anthony Colosimo, Yifeng Li, and Ali Emami. Can we afford the perfect prompt? balancing cost and accuracy with the economical prompting index. *arXiv preprint arXiv:2412.01690*, 2024.

Meta-AI. Llama 3.3 70b instruct model, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. License: Llama 3.3 Community License Agreement, see https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct/blob/main/LICENSE.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*, 2025.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. Openai o3-mini system card, 2025. URL https://openai.com/index/o3-mini-system-card/.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. Inference-time computations for llm reasoning and planning: A benchmark and insights. *arXiv preprint arXiv:2502.12521*, 2025.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/. Dataset licensed under CC BY 4.0.

David Rein. Can good benchmarks contain mistakes? NYU Alignment Research Group Blog, May 2024. Reveals GPQA expert pay ($100/hr) and non-expert solve times. Online: https://wp.nyu.edu/arg/can-good-benchmarks-contain-mistakes/. Accessed 2025-05-14.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof qa benchmark. In *First Conference on Language Modeling*, 2024. Dataset licensed under CC BY 4.0.

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.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*, 2025.

TutorCruncher. Average tutoring rates usa: How much do tutors charge per hour? TutorCruncher Blog, Feb 2025. URL https://tutorcruncher.com/blog/tutoring-business-ideas. Reports $45–$100/hr as typical range for test-prep tutoring. Accessed 2025-05-14.

Upwork. Data entry specialist hourly rates (cost to hire data entry specialist). Upwork Hiring Guide, 2025. URL https://www.upwork.com/hire/data-entry-specialists/cost/. Median $13/hr for data entry freelancers; $10–$20/hr typical range. Accessed 2025-05-14.

U.S. Bureau of Labor Statistics. Customer Service Representatives, August 2025. URL https://www.bls.gov/ooh/office-and-administrative-support/customer-service-representatives.htm.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies. *arXiv preprint arXiv:2406.06461*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*, 2025.

Wyzant Tutoring. New jersey math tutors cost $33 - $55 per hour on average. Wyzant.com (tutoring rate listing), 2025. Average private tutoring rates for math (K-12 and competition). Accessed 2025-05-14.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Zendesk. Average handle time (AHT): Formula and tips for improvement, August 2025. URL https://www.zendesk.com/blog/average-handle-time/.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

A Details of Human Expert Cost Estimation

In this section, we introduce the detailed analysis of how the human expert costs in Table 3 are calculated per dataset.

Dataset	Qualification Requirements	Hourly Rate	Time per Question	Est. Cost
AIME	Advanced high-school contest math skills	\$45–\$100	∼12 minutes	\$9–\$20
BBQ	General familiarity with social biases	\$15	∼0.4 minutes (24 sec)	\$0.10
GPQA Dia.	Graduate-level domain expertise	\$100	∼35 minutes	\$58
GSM8K	Basic arithmetic reasoning	\$33–\$55	∼3.7 minutes	\$2–\$3.50
MATH500	Strong competition-level problem-solving	\$35–\$60	∼12 minutes	\$7–\$12
Two-Digit Add.	Basic numeracy	\$10–\$20	∼0.04 minutes (3 sec)	\$0.01–\$0.02

Table 3: Estimated costs of hiring a human expert to solve one question from each dataset, based on typical qualifications, hourly rates, and time per question.

AIME (American Invitational Mathematics Examination) consists of 15 challenging math problems in a 3-hour contest (administered in two separate sections: `AIME I & II`), giving an average of about 12 minutes per problem (Art of Problem Solving, 2023). In practice, expert math tutors for competitions like `AIME` command high hourly fees in the range of \$45–\$100, reflecting intensive test-preparation rates (TutorCruncher, 2025). This rate range aligns with specialized test prep tutoring in the US, which is higher than regular tutoring due to the advanced problem-solving skills required (TutorCruncher, 2025). At roughly 12 minutes per `AIME` question on average, a solver could handle about five such problems per hour under exam conditions (Art of Problem Solving, 2023).

BBQ (Bias Benchmark for QA) contains short question-answer scenarios targeting social bias. Crowdworkers annotating `BBQ` have been paid around \$15 per hour, a rate chosen to exceed U.S. minimum wage (Parrish et al., 2022). Because each task includes multiple `BBQ` questions, workers were able to answer roughly 5 questions in 2 minutes (Parrish et al., 2022) – i.e. ∼**24 seconds per question**, or about 0.4 minutes per question. This fast per-question time reflects the fact that `BBQ` items are short multiple-choice queries, allowing a human annotator to complete approximately 150 `BBQ` questions in an hour at that pay rate (Parrish et al., 2022).

GPQA-Diamond consists of extremely difficult graduate-level science questions, so human experts demand high compensation. In one case, domain experts were paid about \$100 per hour to contribute and validate `GPQA` questions (Rein et al., 2024). These questions are "Google-proof" and time-consuming: skilled non-expert participants spent **over 30–35 minutes on average per question** when attempting to solve `GPQA` problems with unrestricted web access (Rein et al., 2024). This long duration per question underscores `GPQA`'s complexity: at most 2 questions could be solved in an hour even by motivated annotators, which justifies the premium expert hourly rate (Rein, 2024).

GSM8K contains grade-school level math problems. Solving these is relatively time-efficient for adults: in one study, crowdworkers under time pressure managed to solve about 4.07 `GSM8K` problems in 15 minutes on average (Zhang et al., 2024), or roughly **3.7 minutes per question**. The required skill is comparable to general math tutoring at the K-8 level, for which typical U.S. tutor rates are about \$33–\$55 per hour on platforms like Wyzant (Wyzant Tutoring, 2025). At such a rate, paying a person to solve `GSM8K` problems would be economical, given that a proficient solver can complete approximately 16 questions in one hour (Zhang et al., 2024).

MATH-500 is a set of 500 advanced competition math problems (drawn from the harder tier of a larger MATH dataset). These are similar in difficulty to top-level contest questions such as late `AIME` or Olympiad qualifying problems. As with `AIME`, a well-prepared human might spend on the order of 10–15 minutes per problem, roughly ∼**12 minutes on average for a hard competition question** (Art of Problem Solving, 2023). Tutors capable of solving and teaching such Olympiad-level problems often charge rates on the order of \$50 per hour (with a typical range of \$35–\$60 for competition math tutoring) (Wyzant Tutoring, 2025). Therefore, solving roughly five `MATH-500` problems could cost about \$50 and take around an hour, consistent with the per-question time and high skill required.

Two-Digit Addition consists of simple two-digit addition problems, which are very quick for humans to solve. Early elementary students are often expected to complete about 20-25 basic addition problems in one minute in "mad minute" drills (Fas, 2021). This corresponds to roughly **2–3 seconds per addition** (0.04 minutes per question). Because the task is so elementary, the labor to solve large numbers of such problems can be valued at a lower hourly rate. Simple data-entry style work or basic math tasks on freelance platforms pay on the order of $10–$20 per hour (Upwork, 2025). At $15/hour, for example, a worker could theoretically solve several hundred 2-digit additions within the hour, given the ∼3-second average solution time (Fas, 2021).

Category	Model	Release Date	Cost (per million tokens)	
			Input Tokens	Output Tokens
Lightweight Models	Llama-3.1-8B	7/23/2024	$0.18	$0.18
	GPT-4o Mini	7/18/2024	$0.15	$0.60
	Llama-3.3-70B	12/6/2024	$0.88	$0.88
Large Models	Llama-3.1-405B	7/23/2024	$3.50	$3.50
	GPT-4o	5/13/2024	$2.50	$10.00
	Claude Sonnet-3.5	6/20/2024	$3.00	$15.00
Reasoning Models	OpenAI o1-mini	9/12/2024	$1.10	$4.40
	OpenAI o3-mini	1/31/2025	$1.10	$4.40
	DeepSeek-R1	1/20/2025	$7.00	$7.00
	OpenAI o1	12/5/2024	$15.00	$60.00

Table 4: Per-token inference costs with release dates. Each model name links to the utilized provider.

B Details of Evaluation

For each dataset in our evaluation, we sample up to 128 instances and run each model[†]. $n = 8$ times to estimate the expected runtime cost and accuracy per sample. We use a temperature of 0.7 and `top_p` of 1.0 for all models except OpenAI's reasoning models, for which we set the temperature to 1.0 without applying `top_p`. Additionally, we use the default maximum token generation limits provided by each model. Following Suzgun et al. (2025), we use a concise but descriptive instruction prompt for models to follow:

Experiment Prompt

```
Please solve the following question. You can explain your solution before
presenting the final answer. Format your final answer as:

<answer>
...
</answer>

Instructions:
- For multiple-choice: Give only the letter (e.g., (A)).
- For numeric: Give only the number (e.g., 42).
- For free-response: Provide the full final answer text.

INPUT:

'''
{input}

'''
```

[†]Here, the short-form "model" refers to the underlying model together with its inference pipeline (prompt, decoding settings, etc.). Comparisons throughout the paper are done at this basis, and Appendix B shares the adopted details.

In our experiments, we define the pass $r_m(p)$ as whether the model obtains a correct answer after a single run or not (0 or 1), and the cost $c_m(p)$ as:

$$c_m(p) = n_{\text{in}}(m, p) \cdot c_{\text{in}}(m) + n_{\text{out}}(m, p) \cdot c_{\text{out}}(m) \tag{13}$$

where $n_*(m, p)$ denotes the number of input / output tokens consumed / generated by the model m on problem p, and $c_*(m)$ denotes the dollar costs per input / output tokens consumed / generated by the model m (see Table 4 for the pricing). For the expert costs, we utilize the estimations from Table 3, and set the rates to the upper-bound value to ensure the approximation of the expert accuracy being 1. Finally, as shown in Table 4, we access proprietary models via their original providers, while open-source models are queried through a single provider for consistency and simplicity (TogetherAI, in our case).

C Additional Results

C.1 Expected Accuracy and Inference Costs

As discussed in Section 3.2, we report the expected accuracy and cost for each model per dataset, denoted as $R_m(p \sim D)$ and $C_m(p \sim D)$. To compute these, following the methodology in Section 2.5, we use the i.i.d. sampled set $P \sim D$ of problems per dataset and approximate the expectation by averaging the accuracy $R_m(p)$ and cost $C_m(p)$ across problem instances. The results in Tables 5 and 6 reveal a skewed preference for particular model families under each metric, suggesting that these metrics alone are insufficient to capture the economic impact of models.

Model Category	Basic Quantitative		Knowledge Based		Complex Quantitative	
	2-Digit Add.	GSM8K	BBQ	GPQA Dia.	MATH 500	AIME24
Lightweight Models						
Llama-3.1-8B	89.45	75.78	21.48	17.87	37.30	12.50
GPT-4o mini	99.90	88.57	53.32	18.07	70.02	14.58
Llama-3.3-70B	99.90	92.09	85.06	46.48	72.75	33.33
Large Models						
Llama-3.1-405B	99.71	93.95	85.74	44.14	67.87	31.67
Claude Sonnet-3.5	100	94.43	92.58	55.37	64.75	15.83
GPT-4o	99.71	91.99	90.04	47.07	73.14	14.58
Reasoning Models						
OpenAI o1-mini	99.51	92.58	85.74	49.12	85.94	53.33
OpenAI o1	100	94.04	95.02	73.83	89.45	72.50
DeepSeek-R1	100	93.36	83.69	54.88	93.85	60.83
OpenAI o3-mini	100	92.77	83.79	71.68	88.57	77.08

Table 5: Accuracy (%) per model per dataset: $R_m(p \sim D)$. In each column, the 3 entries with the highest accuracy have blue highlights.

C.2 Evaluation on a Real-World Domain

We evaluate our framework on `Tau-bench` (Yao et al., 2024), a benchmark that targets tool use, agent behavior, and user interaction in real-world domains. We sample 8 tasks per category (airline, retail), totaling 16 tasks, and run each model as an agent under the evaluation protocol described in the original paper. We exclude DeepSeek-R1 due to its visible chain-of-thought being mixed with user messages, which contaminates responses under this protocol. We apply the cost modeling based on total tokens consumed or generated per round, and we aggregate costs over interaction rounds. Estimates are averaged over 4 independent trials per run.

For the human-expert baseline, we consider the "retail or call-center communication" qualification, with an hourly wage of $20.59 (U.S. Bureau of Labor Statistics, 2025) and an average of 6 minutes per task (Zendesk, 2025), which yields $2.06 per task.

Model Category	Basic Quantitative		Knowledge Based		Complex Quantitative	
	2-Digit Add.	GSM8K	BBQ	GPQA Dia.	MATH 500	AIME24
Lightweight Models						
Llama-3.1-8B	4.2e−5	7.4e−5	5.2e−5	1.8e−4	1.5e−4	2.2e−4
GPT-4o mini	5.4e−5	1.9e−4	1.0e−4	3.9e−4	3.7e−4	5.6e−4
Llama-3.3-70B	1.6e−4	3.3e−4	3.1e−4	9.6e−4	6.7e−4	1.1e−3
Large Models						
Llama-3.1-405B	6.9e−4	1.4e−3	1.0e−3	3.0e−3	2.4e−3	3.7e−3
Claude Sonnet-3.5	2.1e−3	3.7e−3	3.0e−3	6.9e−3	5.9e−3	7.5e−3
GPT-4o	2.3e−3	4.5e−3	2.7e−3	0.01	8.7e−3	0.01
Reasoning Models						
OpenAI o1-mini	5.4e−3	8.4e−3	7.6e−3	0.02	0.02	0.07
OpenAI o1	0.02	0.03	0.04	0.25	0.13	0.52
DeepSeek-R1	1.8e−3	5.1e−3	4.6e−3	0.04	0.01	0.04
OpenAI o3-mini	1.1e−3	2.1e−3	2.6e−3	0.01	5.4e−3	0.02

Table 6: Dollar cost incurred per model per dataset: $C_m(p \sim D)$. In each column, the 3 entries with the lowest cost have blue highlights.

Lightweight Models		*Large Models*		*Reasoning Models*	
Llama-3.1-8B	1.6770	Llama-3.1-405B	1.8875	OpenAI o1-mini	1.8230
GPT-4o mini	1.2944	Claude Sonnet-3.5	1.5135	OpenAI o1	1.6406
Llama-3.3-70B	1.6897	GPT-4o	1.2247	OpenAI o3-mini	1.2703

Table 7: Frontier dollar cost-of-pass per model on `Tau-bench` real-world tasks. Each pair of columns lists models (left) and their frontier cost-of-pass with respect to the human expert baseline (right): $V_{p\sim D}(\{m\} \cup \mathcal{M}_0)$. The lowest three values are highlighted in blue, indicating that all the model families have an economic merit in this task.

We repeat the analyses in Sections 3.2, 3.3, and 3.4; and share the results in Tables 7, 8, 9 respectively. Our overall findings indicate that (1) all model families have a merit in this task, (2) the evolution of the frontier cost-of-pass still follows an exponential decay (similar to other tasks), and (3) none of the model families are significantly essential in driving progress.

C.3 Relative Gain per Model Release

Figure 4 presents the relative improvement in temporal frontier cost-of-pass for each model release, illustrated using bar plots. Namely, we calculate:

$$\frac{G_{p\sim D}(\{m_t\}, \mathcal{M}_{t-1})}{V_{p\sim D}(\mathcal{M}_{t-1})} \tag{14}$$

The results indicate that the reasoning models demonstrate notable advancements, particularly on complex quantitative tasks. In contrast, lightweight models exhibit marked gains on basic tasks. These findings support the observations from our experiments (Sections 3.2, 3.4). Notably, The substantial improvement observed for GPT-4o is likely due to it being the first model included in our analysis, resulting in a pronounced leap relative to the baseline cost associated with human expert annotation.

May 13	**Jun 20**	**Jul 18**	**Jul 23**	**Sep 12**	**Dec 5**	**Dec 6**	**Jan 31**
GPT-4o	Claude Sonnet-3.5	GPT-4o mini	Llama-3.1-8B Llama-3.1-405B	OpenAI o1-mini	Llama-3.3 70B	OpenAI o1	OpenAI o3-mini
1.2247	1.1900	0.8411	0.8127	0.8127	0.8021	0.7668	0.7311

Table 8: Frontier dollar cost-of-pass over model release dates on `Tau-bench`. Each column reports the best-to-date frontier value $V_{p\sim D}(\mathcal{M}_t)$ after incorporating models released on the indicated date. The trajectory continues to follow an exponential decay, consistent with other tasks. This table is the tabular version of the time-evolution figure (see Fig. 2 for example).

	Lightweight	Large	Reasoning
Essentialness (%)	22.5	13.2	6.0

Table 9: Essentialness of model families on `Tau-bench` (metric from Section 3.4). The results show that Lightweight models are the most essential, but overall, none of the families are strongly essential in driving progress for the frontier cost-of-pass.

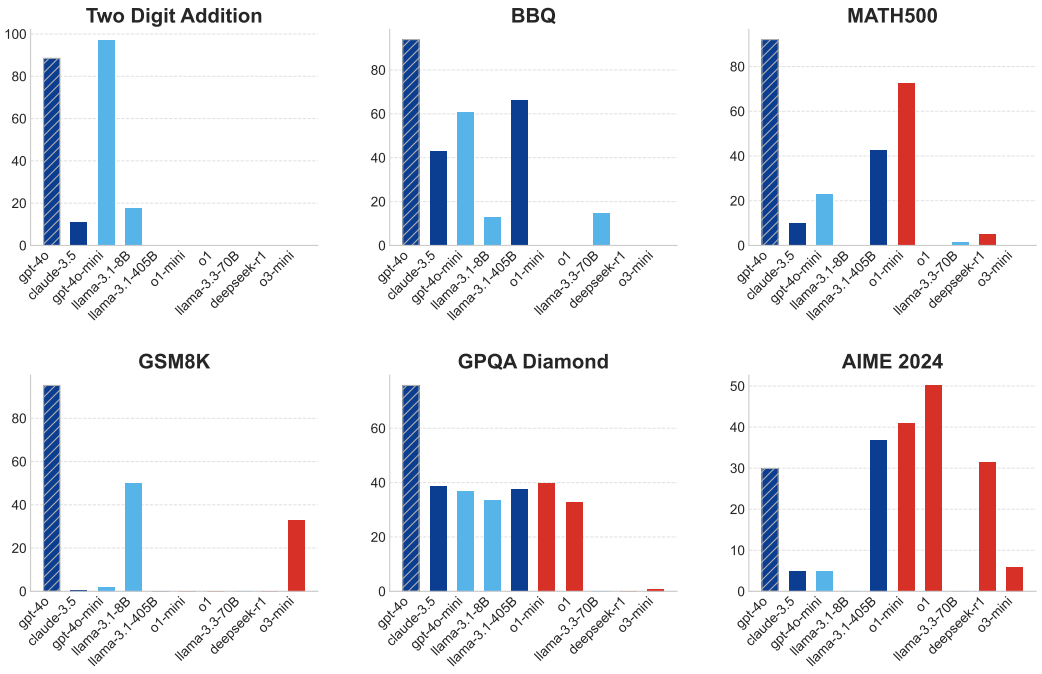

Figure 4: Bar plot showing the percentage of change in frontier cost-of-pass per model release (i.e. $\frac{G_{p\sim D}(\{m_t\},\mathcal{M}_{t-1})}{V_{p\sim D}(\mathcal{M}_{t-1})}$)

C.4 Essentialness of Human Expert Baseline

Adapting the methodology in Section 3.4, we quantify the essentialness of the human-expert baseline for each task. We treat the human-expert baseline as a separate family, $\mathcal{M}_0$, and compare it to the remaining models, $\mathcal{M}_T \setminus \mathcal{M}_0$, via

$$\frac{G_{p\sim D}(\mathcal{M}_0, \mathcal{M}_T \setminus \mathcal{M}_0)}{V_{p\sim D}(\mathcal{M}_T \setminus \mathcal{M}_0)}. \tag{15}$$

Under this definition, essentialness is 100% if there exists at least one instance in the distribution that no model in $\mathcal{M}_T \setminus \mathcal{M}_0$ can solve, so the frontier requires $\mathcal{M}_0$ on some part of the distribution. Conversely, if every instance is solved at strictly lower cost by the LMs, essentialness is 0%.

Applying this analysis, we find that human experts remain fully essential for `GSM8K`, `GPQA-Diamond`, `MATH-500`, and `AIME-2024`, and non-essential (0%) for `BBQ` and `Two-Digit Addition`. Interestingly, there is no task where human experts are partially necessary (in between 0-100%).

C.5 Essentialness of Single Models

In this section, following the methodology outlined in Section 3.4, we quantify the relative improvement in frontier cost-of-pass using a counterfactual approach. Specifically, for each model m_*, we calculate the following:

$$\frac{G_{p\sim D}(\{m_*\},\, \mathcal{M}_T \setminus \{m_*\})}{V_{p\sim D}(\mathcal{M}_T \setminus \{m_*\})}, \tag{16}$$

quantifying the essentialness of the model m_*. The results presented in Figure 5 demonstrate that the contributions of most individual models are largely compensable by the remaining models.

Furthermore, we observe a similar coarse-level trend, as noted in Section 3.4, indicating that different model families provide greater benefits in specific task categories.

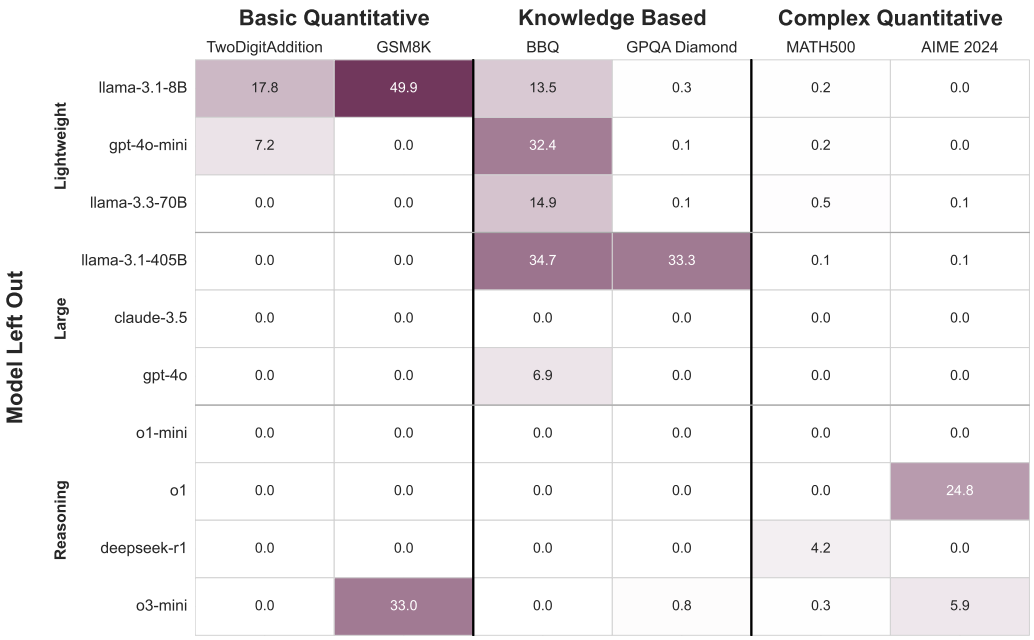

Figure 5: The relative improvement (%) in frontier cost-of-pass under a counterfactual setting, removing a model m_* from the model set $\mathcal{M}_T$. High values mean that the model is essential for maintaining the current frontier.

D Practical Implications, Limitations, and Future Directions

In this section, we acknowledge the limitations of our framework and evaluations, share practical perspectives together with directions for future extensions.

D.1 Extending Modeling Across Complex and Diverse Dimensions

Our experiments consider a common but relatively simple cost and performance modeling, which may not seem clear for practitioners to adapt to their more complex settings. To start with, our analyses use per-token API prices that can be represented by $C_p(m) = \mathbf{w}^\top \mathbf{x}_m(p)$ (Section 2.2), where $\mathbf{w}$ contains prices (input/output tokens) and $\mathbf{x}_m(p)$ contains the corresponding quantities. In practical scenarios, one may include other components of the evaluation pipeline by placing their *unit cost* in $\mathbf{w}$ and their *per-attempt quantity* in $\mathbf{x}$ to enrich the definition. Examples include: verification costs per attempt (e.g. human or automatic checks), tool-usage fees, orchestration overhead (e.g. queue time, cold-start penalties, inter-service latency), and amortized fixed costs per attempt (training, hardware depreciation, maintenance).

Regarding the success metric, one may replace accuracy with a stricter reliability-oriented metric (e.g., passˆk (Yao et al., 2024), requiring k consecutive successes) or a more lenient metric (e.g., pass@k (Chen et al., 2021), rewarding any success within k attempts). Such alternatives are useful in settings where consistency, robustness, or partial correctness matter.

Our evaluations mostly focus on single input/output tasks. More complex settings (multiple turns, tool usage, human verification) can still be handled within our framework by the same extension principle above. We present an instance in Section C.2 on Tau-Bench, where we accumulate consumed and generated tokens across multiple rounds of interaction to fit the framework's setup.

For some applications, alternative units per attempt (FLOPs, time, latency, energy) may matter more than dollar cost, and the application of our framework may not be immediately visible. If an oracle system m' guarantees a non-zero success rate $R_{m'}(p \sim D)$ with a measurable expenditure, one

may treat it as a baseline (analogous to the human expert baseline for costs) and apply our analyses, yielding an alternative unit to dollars. If such an oracle does not exist, assuming that for each $p \sim D$ there exists some $m \in \mathcal{M}$ with $R_m(p) > 0$, several analyses in this paper (e.g. essentialness and impact) still apply.

D.2 Interpretability of Our Framework

Since our cost-of-pass metric is $v(m, p) = \frac{C_m(p)}{R_m(p)}$ for a given problem instance p and model m, with $C_m(p) = \mathbf{w}^\top \mathbf{x}_m(p)$, improvements can arise from (i) lowering unit prices $\mathbf{w}$, (ii) reducing resource use $\mathbf{x}_m(p)$, or (iii) increasing success probability $R_m(p)$. In practice, *cheaper tokens* reduce $\mathbf{w}$ via pricing changes, distillation, quantization, or optimized serving; *fewer tokens* reduce $\mathbf{x}_m(p)$ via prompt compaction, dynamic budgets, or instructions that promote concise generation; and *higher accuracy* increases $R_m(p)$ through better prompting, light test-time techniques, or improved model / training. Thus, the metric and framework capture these practical dimensions and quantify them in an interpretable way.

While these dimensions explain directional changes, the formulation still only reports expectations and therefore does not incorporate variance. Two strategies with identical expected cost-of-pass may entail very different variances, and hence different risks. Augmenting the metric with variance or risk-adjusted objectives would enhance interpretability and practical usefulness, and left for future work.

D.3 Limitations

We present limitations associated with both the framework and our evaluations. While covered in Section D.1, our evaluations instantiate simple formulations for costs and success. This is a reasonable proxy from a user perspective and extends gracefully, but it still omits indirect and context-specific terms (like evaluation/verification overheads, wait times, invocation retries, tool-call charges etc.). Our framework remains compatible with these terms via the vectorized cost view, but we do not include them in our core results. Regarding the success metric, our framework assumes a binary success/failure criterion, thus continuous or composite notions of success are not modeled directly.

Both pricing and performance can vary across API providers (Gao et al., 2025), especially for open-source models hosted by third parties. Treating each provider–model (ie. inference pipeline) pair as a distinct strategy and either (i) reporting all results from the same provider consistently or (ii) providing multiple provider snapshots per model can make benchmarking and comparisons more robust.

Throughout our evaluations, we fix a single concise instruction and sampling arguments (e.g., temperature, top-p). We chose this to reduce degrees of freedom and enable comparability across models. However, results may be sensitive to these choices. Future work can study prompt and decoding sensitivity by evaluating small prompt ensembles per model and conducting sweeps over decoding settings.

Model selection in our evaluations can introduce temporal and categorical bias. Due to budget, compactness, and coverage considerations; we evaluated a subset of releases. For this, we fixed a short time window and chose representative models per major family to capture broad trends. A more exhaustive design is beyond our scope, but two extensions are natural: (i) broadening coverage to include historical and subsequent releases, and (ii) sampling more densely within a fixed horizon (more models at closely spaced release dates).

Our family distinction between lightweight and large models is based on per-token prices. Alternative categorizations (parameter count, open/closed status, deployment modality) are possible. We prioritize transparency and reproducibility; as sizes are often undisclosed, and openness does not map directly to user-incurred costs. We also keep the analysis prototypical by focusing on user-facing, common-case models (omitting quantizations or distillations). Future work can adopt alternative categorizations to quantify economic impact under different groupings.

The human expert baseline assumes that qualified annotators always succeed given sufficient time and compensation. Extremely challenging problems (or scarce expertise) may violate this assumption.

Rigorous human subject studies could estimate a "human cost-of-pass," capturing both success probability and variance.

Despite these caveats, the framework's abstract, modular design means each of the above extensions can be implemented by plugging in refined cost functions, richer success metrics, or additional variability terms. At the same time, our core analysis remains a practical baseline, as per-token API pricing reflects actual user-side costs, and binary pass/fail captures minimal utility in many applications. We hope future work adapts the framework along these lines and develops datasets that jointly stress cost and performance dimensions.