The Latent Role of Open Models in the AI Economy*

Frank Nagle[1] and Daniel Yue[2]

[1]Massachusetts Institute of Technology & The Linux Foundation
[2]Georgia Institute of Technology

Draft Date: November 18, 2025

Abstract

The rapid diffusion of large language models (LLMs) is mediated by an emerging market for AI inference. However, its economic structure is poorly understood due to challenges in measuring LLM usage. While some fear AI will inevitably evolve into an oligopolistic structure with winner-take-most dynamics, others argue that the rapid emergence of capable open models will inevitably commoditize AI inference. Leveraging data on model-level usage and prices from OpenRouter comprising just under 1% of the total market for LLM inference, we empirically uncover the latent role of open models relative to closed models. Closed models dominate, with on average 80% of monthly LLM tokens using closed models despite much higher prices — on average 6x the price of open models — and only modest performance advantages. Frontier open models typically reach performance parity with frontier closed models within months, suggesting relatively fast convergence. Nevertheless, users continue to select closed models even when open alternatives are cheaper and offer superior performance. This systematic underutilization is economically significant: reallocating demand from observably dominated closed models to superior open models would reduce average prices by over 70% and, when extrapolated to the total market, generate an estimated $24.8 billion in additional consumer savings across 2025. These results suggest that closed model dominance reflects powerful drivers beyond model capabilities and price — whether switching costs, brand loyalty, or information frictions — with the economic magnitude of these hidden factors proving far larger than previously recognized, reframing open models as a largely latent, but high-potential, source of value in the AI economy.

*Authors listed alphabetically. Correspondence to: daniel.yue@scheller.gatech.edu. We thank seminar participants at MIT FutureTech and Georgia Tech Scheller for thoughtful comments. We thank Zahra Rasouli for assistance in constructing the model crosswalks. We thank AI models (both open and closed) for excellent research assistance. All errors remain our own.

1 Introduction

Artificial intelligence (AI) is a general purpose technology reshaping fundamental economic processes from basic science to production lines and everything in between, with the potential to fundamentally reorganize the economy (Eloundou et al., 2024; Jones, 2025). In particular, large language models (LLMs)[1], have demonstrated high performance on many economically relevant tasks, spurring adoption of chatbots at a record pace and driving increasingly sophisticated applications (including AI agents). These applications are built on top of LLMs via application programming interfaces (APIs) that allow direct, programmatic access to LLM functionality. This promise has catalyzed historic levels of investment in AI infrastructure, with an estimated $520 billion in capital expenditure—over 1% of U.S. GDP—anticipated in 2025 alone (Kedrosky, 2025). Such unprecedented investment, alongside widespread concerns about industry concentration (Acemoglu and Johnson, 2023; Barr, 2024; Milmo, 2024; Ramzanali, 2025), reflects an oligopoly narrative where leading firms can recoup the massive fixed costs associated with infrastructure and model training through sustained market power and pricing advantages. Yet this view stands in stark tension with a narrative that AI is a commodity, most famously exemplified in a leaked internal memo from Google entitled "We Have No Moat, and Neither Does OpenAI" (Patel and Ahmad, 2023). The memo warned that while closed-model providers have been competing with each other, open models have been "quietly eating our lunch," suggesting that AI capabilities can be rapidly mimicked, potentially eroding market power and driving companies to seek competitive advantage at other levels of the AI stack.

At the core of this tension lies a critical dichotomy between open and closed models. Closed models—such as those from OpenAI, Anthropic, and Google—keep model details proprietary and require users to pay for both the model and the underlying compute to access inference services[2] (using a pre-trained foundation model to generate output from the user's prompt). In contrast, open models[3]—such as those from Meta (Llama), IBM (Granite), DeepSeek, and Mistral—make certain model details public (weights, source code, architecture, or training data), allowing any company to host the model (i.e., become an "inference provider") or users to run it locally. This openness enables inference at only the cost of compute power, essentially making the software free and creating competitive pressure akin to commodity markets. If we want to understand the

[1] LLMs are generative AI models that are trained on massive datasets and are then used to create new content based on the patterns learned from the training data.

[2] In AI, "inference" refers to using a pre-trained model to generate predictions or outputs, whereas in economics and statistics, "inference" typically refers to estimating parameters from data—what AI practitioners would call "training."

[3] There has been a long-running debate since the Open Source Initiative (OSI) released their definition of open source AI, which requires the model to have open weights, open source code, open architecture, and open data to earn the open source moniker (Open Source Initiative, 2024). Many have argued that even models that are only open weight (the weights for the neural network underlying the model are made open) still create substantial value through their openness since users can freely implement the model even if they cannot see the underlying source code or training data. In this paper, we utilize this broader view of openness, as even an open weight model can be used for free, thus putting downward price pressures on other models and producing the "direct effect" that we study in this paper.

market structure for LLM inference, then we must first understand the relationship between open and closed models. Therefore, we ask: "what role do open models play in the AI economy?"

Answering this question is difficult due to challenges in measuring LLM usage. If measuring closed model usage is hard due to the tightly guarded nature of siloed model provider usage data, measuring open model usage completely is nearly impossible because they may be used locally in a way that is challenging to measure in any convincing centralized manner. As a first step to solving these measurement challenges, we utilize a data source that provides insight into the LLM inference market: OpenRouter[4], a platform for LLM inference services that connects users (frequently companies or AI applications) to AI inference providers. OpenRouter describes itself as "The Unified Interface for LLMs," allowing users to engage with many inference providers (including both closed model providers such as OpenAI or Anthropic, and third-party inference providers that offer multiple open models) through one interface rather than having to setup individual integrations with each inference provider. Crucially, by studying open model inference providers, we can measure some priced, non-local subset of open model usage, enabling estimates of the marginal cost of inference for such models. Further, OpenRouter's data cuts across all prominent vendors and models (both closed and open), enabling comparative analysis. We are not the first to use OpenRouter to study the economics of LLMs: Fradkin (2025) uses this setting to study substitution and competition among market-leading (closed) models[5]. However, by combining this unique dataset with other data on model performance benchmarks from Artificial Analysis and LMArena, we produce one of the first large-scale analyses of usage and pricing across the market for LLMs that reveals the latent role of open models.

Our findings reveal a puzzle in the current state of the LLM inference market through three key observations. First, closed models dominate the LLM inference economy—accounting for, on average, 80% of token usage and over 95% of revenue passing through OpenRouter. Second, this dominance is not driven by a substantial performance gap: while closed models maintain a small ($\sim$ 10%) performance advantage on key benchmarks, open models consistently catch up within 3-6 months, with this catch-up time accelerating over our observation period. Third, open models benefit from significantly lower prices (approximately 84% lower than closed model prices on a usage-weighted basis), driven by competition among the many third-party inference providers that can offer these models.

If open models offer comparable performance at substantially lower prices, why do closed models continue to dominate? This pattern suggests significant underutilization of open models, but such underutilization need not imply irrational consumer behavior[6]. Rather, it indicates that any model of consumer choice where AI model choice decisions are driven solely by observable price and capability benchmarks is incomplete. Additional factors beyond price and benchmark

[4]https://openrouter.ai/

[5]Since the Fradkin (2025) working paper release, the author has expanded their work into a not-yet-public broader study of supply and demand in the market for LLMs (Demirer et al., 2025). We'd like to acknowledge their ongoing work as related and complementary to our study.

[6]We refer to users of LLM inference as 'consumers' to reflect their role as purchasers, but recognize that many of these consumers are in fact firms or downstream products rather than individuals.

performance—such as switching costs, brand/trust, or aversion to foreign models[7]— must be influencing model selection. We frame our estimation exercise as quantifying the implicit value that consumers place on these unobserved factors. By identifying instances where users select closed models that are observably dominated[8] (that are worse performing and more expensive) than open alternatives, we can estimate the foregone consumer savings. Through a counterfactual simulation where dominated closed models switch to superior open alternatives, we estimate that this underutilization represents $104-$146 million in 2025 on OpenRouter alone—approximately 57-80% of the platform's total revenue depending on matching strategy. To extrapolate to the broader LLM inference market, we employ three independent approaches using different reference values: Menlo Ventures' market survey data ($35.1B market estimate for 2025), publicly disclosed Google token volumes ($42.7B estimate), and estimated OpenAI API revenue ($60.4B estimate). Across these methods, our estimates suggest potential unrealized consumer savings ranging from $20.1 billion to $48.3 billion annually, with our preferred estimate of $24.8 billion per year based on the Menlo Ventures approach with median underutilization rates. This represents roughly 70% of total market spending. For comparison, a complementary analysis examining the reverse scenario—swapping current open model users to closed alternatives—yields substantially smaller value estimates of $350 million to $1.23 billion, more than an order of magnitude less—indicating that the realized value of open models is much smaller than the unrealized value. These surprisingly large magnitudes underscore the economic significance of understanding what drives model selection beyond the traditional price and performance considerations.

These findings contribute to two distinct but complementary literatures. First, we contribute to the emerging literature on market structure and strategy for LLMs. While recent theoretical work has developed formal models of incentives for open model production and their market effects (Azoulay et al., 2025; Habibi, 2025; Leisten, 2025; Xu et al., 2024), empirical evidence remains scarce. Fradkin (2025) uses OpenRouter data to study competition among closed-source market leaders, but ours is among the first to empirically characterize the relative role of open and closed models, demonstrating that closed model dominance stems from factors beyond pure capability advantage served at competitive prices. Second, we extend the literature on the economic value of open source software (Blind et al., 2021; Greenstein and Nagle, 2014; Hoffmann et al., 2024;

[7]We organize the many possible factors that could explain this underutilization in the Discussion section, but are unable to disentangle them empirically in this paper.

[8]Throughout this paper, we use terminology inspired by concepts from welfare economics—particularly the notions of dominance and consumer savings (inspired by the formal welfare analysis concepts of Pareto domination and consumer surplus). However, we emphasize an important limitation: our analysis does not account for the fact that consumers may optimize their model choices based on many features beyond the observable price and performance metrics we measure. We can only describe domination with respect to these observed dimensions—price per token and benchmark scores—without accounting for numerous other model characteristics that may be important to users. We believe this approach remains valuable because the prevailing mental model in the AI industry—exemplified by the widely-circulated memo "We Have No Moat, and Neither Does OpenAI"—suggests that price and performance should largely determine model choices in a commodity-like market, while other factors are presumed transient in their effect. The substantial underutilization we document relative to this simple model is therefore arguably interesting and economically significant. We discuss additional variables and considerations that may be important for accurately describing consumer choice in the Discussion.

Korkmaz et al., 2024; Nagle, 2019) to the domain of open AI models. This literature has also long faced measurement challenges due to the absence of market prices for open source software (Hoffmann et al., 2024). Additionally, recent work by Collis and Brynjolfsson (2025) estimates $97 billion in consumer surplus from AI chatbots like ChatGPT using survey methodology, a related but complementary product market. Our setting is unique in featuring direct competition between open and closed alternatives at comparable capability levels, allowing us to quantify the unrealized value of open models. We note an important caveat: our estimates capture only the direct usage effect of open models (cost savings from substitution), not their broader value through knowledge spillovers or as foundations for fine-tuning, which may be substantial (possibly even larger than the value we measure here) but have yet to be empirical quantified. Despite this limitation, the large magnitude of unrealized direct-use value we uncover suggests that understanding the full economic impact of open models remains an important agenda for future research.

2 Setting and Methods

2.1 The Market for LLM Inference

To understand the role of open models in the AI economy, we must first introduce the market structure for large language model inference. Figure 1 illustrates the key participants and relationships in this market, which we describe in detail below.

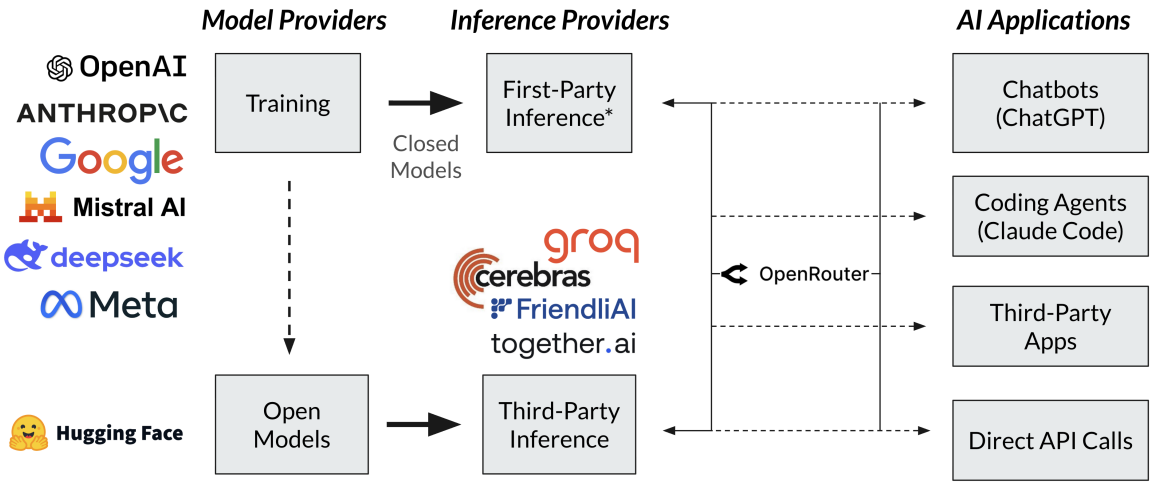

Figure 1: The Market for LLM Inference. The * is added next to First-Party Inference providers to indicate that they may also provide inference services through commercial cloud partnerships (such as OpenAI with Microsoft Azure or Anthropic with Google Cloud). In this case, such a service would be second-party inference, served alongside first-party.

The production of large language models begins with model providers who undertake the substantial fixed costs of model training. Training a frontier LLM requires careful architectural design,

curation of massive datasets spanning billions or trillions of tokens, and access to thousands of high-performance GPUs operating continuously for weeks or months. The result of this training process is a model consisting of billions of numerical parameters (weights) that encode patterns learned from the training data. After completing the training phase, model providers face a critical strategic decision: whether to keep their models proprietary or to release them as open models. Closed models maintain their weights, architecture, training data, and other details as trade secrets, while open models make some combination of these elements publicly available. In practice, most prominent open models are hosted on Hugging Face[9], which has emerged as the de facto repository for open AI models, making them accessible to anyone who wishes to download and deploy them.

For model providers who choose to keep their models closed, the natural next step is to serve these models directly to end users through first-party inference services[10]. Inference refers to the process of using a trained model to generate outputs in response to user prompts, and it represents the primary mechanism through which model providers can monetize their substantial training investments, with pricing often given in units of dollars per million tokens for each of prompt and output tokens[11]. Some closed model providers choose to vertically integrate further by building consumer-facing applications on top of their inference services. OpenAI's ChatGPT represents perhaps the most prominent example of this vertical integration strategy, layering a conversational interface and "agentic" tool integration directly onto the company's GPT models. Nevertheless, most inference providers also offer API access to their models, enabling other organizations to build applications that utilize their language models programmatically.

Since the release of GPT-3.5 through OpenAI's API in March 2023, a significant area of innovation in the AI economy has been the development of complementary applications built on top of large language models through these APIs. These applications extend well beyond chatbots, and include the recent proliferation of software coding agents such as Claude Code or Cursor. Third-party developers, especially other companies, have created specialized applications ranging from external services like custom customer service agents, or internal-facing services like documentation search systems, alongside direct API integrations useful for research and automation purposes. Such applications may be sold as products or services on other downstream markets, or as value-added features layered on existing products or services. As the AI economy continues to mature, these application-layer innovations enabled by LLM APIs will likely proliferate further, and a substantial portion of economic value creation (and likely value capture) will occur at this layer of the stack rather than at the model inference layer itself.

When a model is released as open, it becomes available for any party to download and deploy

[9]Other popular platforms for accessing and deploying open models include Ollama (Ollama Contributors, 2024), which streamlines local LLM deployment, and GPT4All (Nomic AI, 2024), which provides user-friendly interfaces for running models locally.

[10]As noted in the caption, model providers may also serve the model through commercial partners, a form of "second-party inference," although we conceptually group these forms of inference provision in this work because they do not constitute competitors to the first-party.

[11]For example, on November 11th, 2025, OpenAI's GPT-5 cost $1.25 / 1M token for input tokens, and $10.00 / 1M for output (response) tokens.

independently. This openness has enabled the emergence of a distinct set of players in the market: third-party inference providers such as Groq (Groq, Inc., 2024), Cerebras (Cerebras Systems, 2024), FriendliAI (FriendliAI, 2024), and Together AI (Together AI, 2024). These companies specialize in optimizing the serving infrastructure for open models, differentiating themselves primarily along dimensions of latency, time to first token, and price. Much of their technical work involves developing efficient hardware kernels and systems-level optimizations to deliver large language model inference more cost-effectively than organizations could achieve by deploying models themselves[12]. This competitive dynamic among multiple providers serving the same open models creates downward pressure on inference prices, as we document empirically in our results section.

The result of these dynamics is a market for LLM inference characterized by both first-party providers (who train and serve their own models) and third-party providers (who serve open models trained by others). Each category of provider maintains its own API, pricing structure, and performance characteristics. This market structure creates both opportunities and challenges for developers seeking to integrate LLMs into their applications, as they must navigate across multiple incompatible APIs and evaluate tradeoffs between cost, model capabilities, and service quality.

It is important to emphasize several caveats regarding the scope of our analysis. Most notably we do not focus on vertically integrated inference activities like those supporting chatbots which frequently operate under a subscription business model and constitute a significant portion of the AI economy today. Further, we focus specifically on text-based LLM inference and do not attempt to capture all AI inference activities. In particular, we exclude generative media applications like image and video generation (such as those provided by fal.ai (Fal AI, 2024)), which may represent substantial economic activity but operate in distinct markets with different dynamics. We also exclude inference for fine-tuned models, whether fully fine-tuned or using parameter-efficient methods such as LoRAs[13], as these represent customized deployments that do not participate in the standardized inference market we study. Additionally, while we focus on open models as a key source of openness in the AI stack, it is worth noting that open source software pervades every layer of this ecosystem. Model training relies heavily on open source framework ecosystems such as PyTorch (Paszke et al., 2019; Yue and Nagle, 2025) and Transformers (Wolf et al., 2020). Model serving (including for inference) builds on open source tools like vLLM (Kwon et al., 2023) and ONNX (ONNX Contributors, 2024). Application development leverages open source projects such as AutoGPT (Richards and Significant Gravitas Contributors, 2023), LangChain (Chase and LangChain Contributors, 2024), and Open Web UI (Open WebUI Contributors, 2024). Therefore, while open models represent the most visible and perhaps economically significant layer of openness

[12]It is worth noting that much of this inference infrastructure itself builds on open source software, particularly vLLM (Kwon et al., 2023), meaning that any competitive differentiation among inference providers must occur through optimizations layered on top of these shared open source foundations.

[13]Fine-tuning refers to the process of further training a pre-trained language model on a specific dataset or task to adapt it for particular applications. Full fine-tuning updates all model parameters, which can be computationally expensive and require substantial GPU resources. LoRA (Low-Rank Adaptation) represents a parameter-efficient alternative that freezes the original model weights and introduces trainable low-rank matrices into specific layers, allowing effective adaptation while updating only a small fraction of parameters—typically reducing trainable parameters by orders of magnitude while maintaining comparable performance to full fine-tuning.

in the AI stack, the broader ecosystem depends on substantial open source contributions across all layers. A comprehensive accounting of the economic value created by openness in AI would need to address all of these contributions, though such an analysis lies beyond the scope of this paper.

2.2 Data

2.2.1 Market Data from OpenRouter

Given the fragmented market structure described above, measuring usage patterns across both open and closed models presents a significant empirical challenge. Direct data from first-party providers like OpenAI or Anthropic would reveal usage of their closed models but would not capture instances where their users also experiment with or deploy open alternatives. Conversely, data from third-party inference providers like Groq or Together AI would have open model usage but would miss closed model consumption by those same users. This measurement challenge motivates our use of data from OpenRouter, a platform that provides a unified interface across both closed and open model providers, who's position is visually depicted in Figure 1.

OpenRouter operates as an inference aggregation platform that routes API requests from applications to any of dozens of underlying inference providers, supporting both closed models (served by their original creators or commercial partners) and open models (served by specialized third-party inference providers). This routing functionality lowers switching costs and enables interoperability across the fragmented inference market.[14]. From a research perspective, OpenRouter provides the optimal perspective for studying comparative usage of open and closed models precisely because it aggregates demand across both categories.

OpenRouter launched in early 2023 and has grown rapidly to become the primary routing platform in the AI inference market. As of November 2025, the platform serves over 4 million developers across more than 500 AI models from over 60 inference provider integrations. In its most recent funding round, OpenRouter raised $40 million from prominent venture capital firms including Andreessen Horowitz, Menlo Ventures, and Sequoia Capital (OpenRouter, 2025). The platform operates on a straightforward business model: it applies a 5.5% markup[15] to the credits purchased on the platform, then passes costs directly through from underlying inference providers. This markup structure ensures OpenRouter has strong incentives to maintain accurate, up-to-date pricing information, as outdated prices would directly impact their ability to route requests and earn revenue—a valuable feature that distinguishes this data source from other decentralized datasets of LLM prices, which may contain stale or inaccurate pricing information. This pricing transparency, combined with the platform's scale, makes OpenRouter the dominant player in the inference routing space, far outpacing competitors such as Eden AI or Portkey.

[14]By default, OpenRouter distributes requests among multiple inference providers for the same model using price-based load balancing (documentation) Consequently, our analysis focuses on model-level competition rather than within-model provider-level competition, as we do not have access to inference provider-specific usage data, and these choices may be driven by default algorithms rather than intentional customer choice.

[15]This OpenRouter added "tax" rate is accurate as of November 10, 2025.

A critical consideration in using OpenRouter data is selection bias. The platform's users represent a selected subsample of the broader population of LLM consumers, and this selection could bias our estimates of relative open versus closed model usage in either direction. Several factors suggest we may overrepresent open model usage among the users we do observe. OpenRouter users have specifically sought out a platform emphasizing interoperability and exploration across multiple providers, suggesting higher propensity to experiment with alternatives to default closed model providers. The platform explicitly lowers switching costs when moving away from providers like OpenAI and Anthropic, which may increase open model usage relative to settings where switching requires more effort. Finally, while its true that we underrepresent open model usage because we cannot observe local deployments by sophisticated individuals and organizations, we also underrepresent closed model usage from those same organizations when concerned with security, as they contract directly with closed model providers rather than using an aggregation platform. In aggregate, we believe that this selection bias concern cuts in support of our main empirical findings rather than against them. Our central result is that open models are underutilized relative to what their price-performance characteristics would predict. If OpenRouter users are actually more inclined toward open models than the general population, then measuring underutilization in this selected sample suggests that underutilization in the broader market is even more pronounced. The bias works against finding our main result, making our estimates conservative lower bounds.

Importantly, even if local deployment of open models is substantial and our platform-based measurement significantly underestimates total open model usage, OpenRouter still provides valuable insight into the marginal cost of serving these models through specialized providers. This pricing information remains informative for understanding the cost differential between open and closed models even if usage volumes are not fully representative.

We collect daily model-level data from OpenRouter covering the period from May 2025 through September 2025. For each model available on the platform, we observe total daily token usage (separately for prompt and completion tokens), the set of inference providers offering that model, and the price per million for each type of token charged by each provider. When multiple providers offer the same model at different prices, we use the minimum non-zero price in our analysis, as users can always choose the lowest-priced provider[16]. When imputing revenue we multiply prompt tokens with prompt price and completion tokens with completion price and sum the values, but when we analyze price alone we simply use the prompt token price because prompt tokens tend to be much more numerous than output tokens. We also collect model metadata including creation dates, provider information, and associated Hugging Face repository links, enabling us to determine which models are open. This data structure enables us to analyze usage patterns, pricing dynamics, and competitive dynamics across both open and closed models within a unified framework.

[16]Our results are robust to using the median price across providers instead of the minimum. The key intuition is that variation in prices across providers for the same model is relatively small compared to the substantial price differences between open and closed models.

2.2.2 Benchmarks from Artificial Analysis and LMArena

To complement our usage and pricing data from OpenRouter, we merge in model performance benchmarks from two leading evaluation platforms: Artificial Analysis and LM Arena. These benchmarks enable us to assess whether usage patterns align with objective measures of model capabilities, or whether other factors drive model selection.

Artificial Analysis[17] operates as an independent platform providing comprehensive evaluation and comparison of AI models across multiple performance dimensions. Importantly, the platform conducts its own assessments rather than relying solely on vendor-reported scores, maintaining both lab-claimed performance metrics from model creators and independently computed results (Artificial Analysis, 2025). We utilize Artificial Analysis data for several key benchmarks that span different capability domains. MMLU Pro (Wang et al., 2024) represents an enhanced version of the Massive Multitask Language Understanding benchmark, testing knowledge across 14 domains including biology, business, chemistry, computer science, economics, engineering, health, history, law, math, philosophy, physics, and psychology. GPQA (Rein et al., 2024) (Graduate-Level Problems in Question Answering) evaluates models on PhD-level science questions, providing a measure of reasoning capability on specialized technical content. LiveCodeBench (Jain et al., 2024)[18] measures code generation abilities.

LM Arena[19] (Chiang et al., 2024), formerly known as Chatbot Arena, provides a complementary assessment methodology based on crowdsourced human preferences rather than automated evaluation metrics. Operated by the Large Model Systems Organization (LMSYS), the platform collects blind comparisons where users interact simultaneously with two anonymous models, then vote for which response they prefer. After aggregation, the resulting ratings reflect each model's performance on open-ended dialogue and instruction-following tasks. LM Arena ratings provide a particularly valuable signal for real-world usefulness because they derive from human preferences on naturally occurring tasks rather than curated test sets. We access these ratings through the platform's public leaderboard via the LMArena HuggingFace Space repository.

Linking benchmark data to OpenRouter usage data requires matching models across platforms despite inconsistent naming conventions. For instance, the same underlying model might appear as `meta-llama/Llama-3.1-70B-Instruct` on Hugging Face, `meta-llama-3.1-70b-instruct` on OpenRouter, `llama-3.1-70b-instruct` on Artificial Analysis, and `Llama 3.1 70B` on LM Arena. Version identifiers, quantization formats, and instruction-tuning variants further complicate matching. To address this challenge, we implement a multi-stage matching process combining exact matching with fuzzy string similarity methods. We begin by applying a curated set of hardcoded mappings for known difficult cases, developed iteratively through manual review. We then attempt exact matching on standardized model names (lowercased and whitespace-normalized), followed

[17] `https://artificialanalysis.ai/`

[18] LiveCodeBench focuses on algorithmic programming problems from coding competitions (LeetCode, AtCoder, CodeForces). This differs from another popular coding benchmark, SWE-bench (Jimenez et al., 2024), which evaluates repository-level software engineering tasks using real GitHub issues and pull requests.

[19] `https://lmarena.ai/`

by slug-based matching for platforms providing structured identifiers. For remaining unmatched models, we use substring containment checks and Levenshtein distance calculations to identify near-matches differing by minor variations. Each automated match is reviewed for accuracy, with particular attention to high-usage models that will substantially influence our analyses. Appendix A.1 provides complete technical details on our crosswalk methodology.

3 Results

We organize our results into two groups that together reveal a puzzle in the LLM inference market. The first group establishes an empirical pattern: 1) closed models dominate the LLM inference economy despite higher prices, 2) open models catch up to the capabilities of closed models within a few months, and 3) the existence of open models enables competition, driving down prices. Given that open models offer comparable performance at dramatically lower prices yet closed models dominate, we turn in the second group of results to quantifying the economic significance of this apparent market inefficiency. We document that open models are substantially underutilized by consumers relative to what their price and performance characteristics would predict.

3.1 Closed Models Dominate the LLM Inference Economy Despite Higher Prices and Modest Performance Benefits

3.1.1 Closed Models Dominate the LLM Inference Economy

There are many ways to measure economic activity at the company level, but perhaps the most prominent is revenue. Therefore, to illustrate the raw data, Figure 2 shows imputed weekly revenue on OpenRouter from inference providers aggregated to the model provider level in dollars. Using the log scale helps with visualization but masks the dominance of Anthropic's models over this period. For example, in the week of September 14, while Anthropic had $2.70MM in revenue via OpenRouter, Google had less than half of that ($831K), and OpenAI and xAI, in turn, had roughly half of that ($359K and $348K respectively). The leading revenue generator for open models (DeepSeek) had roughly 2.8% of Anthropic's revenue ($76K). Aggregating the closed and open models from the top 10 providers shows that closed models brought in $4.2MM in revenue that week through OpenRouter, while open models brought in $142K, or 96% less revenue. Figure A2 in the appendix emphasizes these substantial differences by showing the same data on a linear scale, highlighting the dominance of closed model providers even more clearly.

To understand the aggregate patterns more clearly, Figure 3 presents four key metrics comparing open and closed models across our analysis period. Panel A shows that open models vary from 15-30% of token share, averaging 20.67% of all tokens processed via OpenRouter per week over the analysis period. That is, open models constitute a substantial, but minority usage. Panel B further reveals that these same open models account for only 4.18% of total revenue on average. These modest descriptive results quantifying open model's usage share (20%) and revenue share (4%) demonstrate that closed models dominate the LLM inference economy on OpenRouter. These

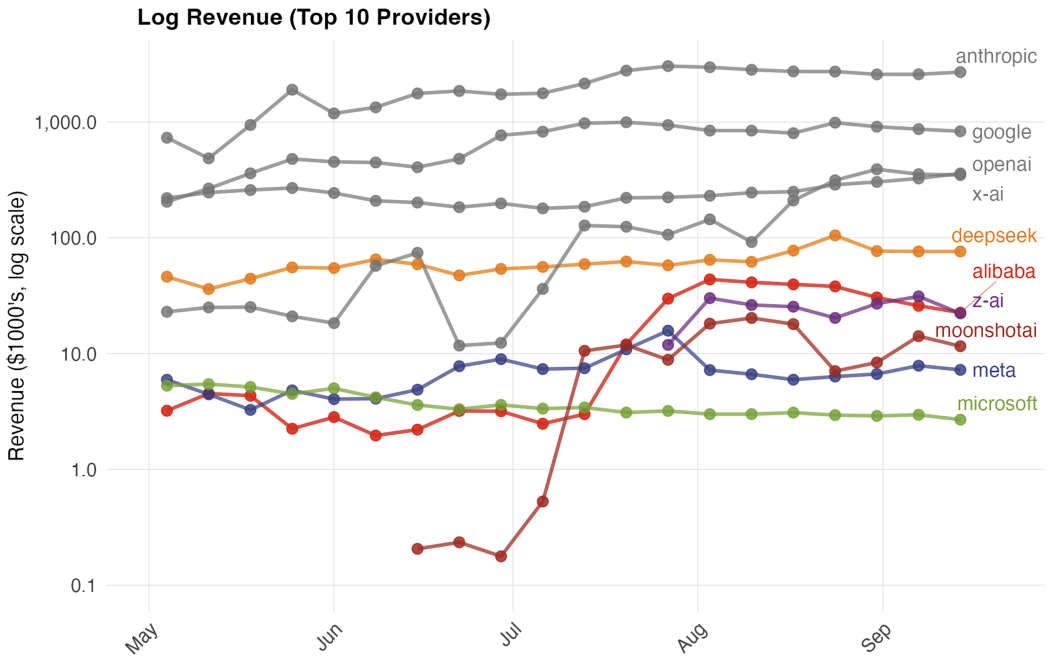

Figure 2: Weekly Revenue by Model Provider (Log Scale). Weekly revenue aggregated by model provider from May to September 2025, showing the top 10 providers on a logarithmic scale. Open models are shown in color, closed models are grey.

results therefore constitute the main finding of this section, and we seek to unpack them throughout our remaining analysis.

Why the gap between usage and pricing rates? The answer becomes clear when examining pricing and performance. Panel C shows that, weighted by usage[20], open models average just 15.66% of the cost of closed models—in other words, open models are approximately six times cheaper than their closed counterparts. Indeed, prices remained relatively fixed within a model family over this time period, meaning that the price gap between Open and Closed models has been relatively fixed as well (see Figure A4). Yet Panel D demonstrates that this price difference is not driven by a large commensurate capability gap: open models average 89.6% of closed model performance on GPQA, a widely used benchmarking metric for PhD-level reasoning tasks. Similar patterns recur for other benchmarks. Thus, open models deliver nearly 90% of the capability at roughly 16% of the price, yet capture only 4% of revenue. The consistency of these high-level, aggregated patterns motivates a deeper dive into each pattern in subsequent analysis.

[20]Throughout this paper, "weighted by usage" refers to computing averages where each model's contribution is proportional to its token usage. Formally, for a set of models with prices p_i and token counts t_i, the usage-weighted average price is $\bar{p} = \sum_i (p_i \cdot t_i) / \sum_i t_i$. We also apply this weighting to benchmark capabilities when aggregating to a set of models (open, closed, or all). This weighting ensures that heavily used models have proportionally greater influence on the aggregate statistics than rarely used models.

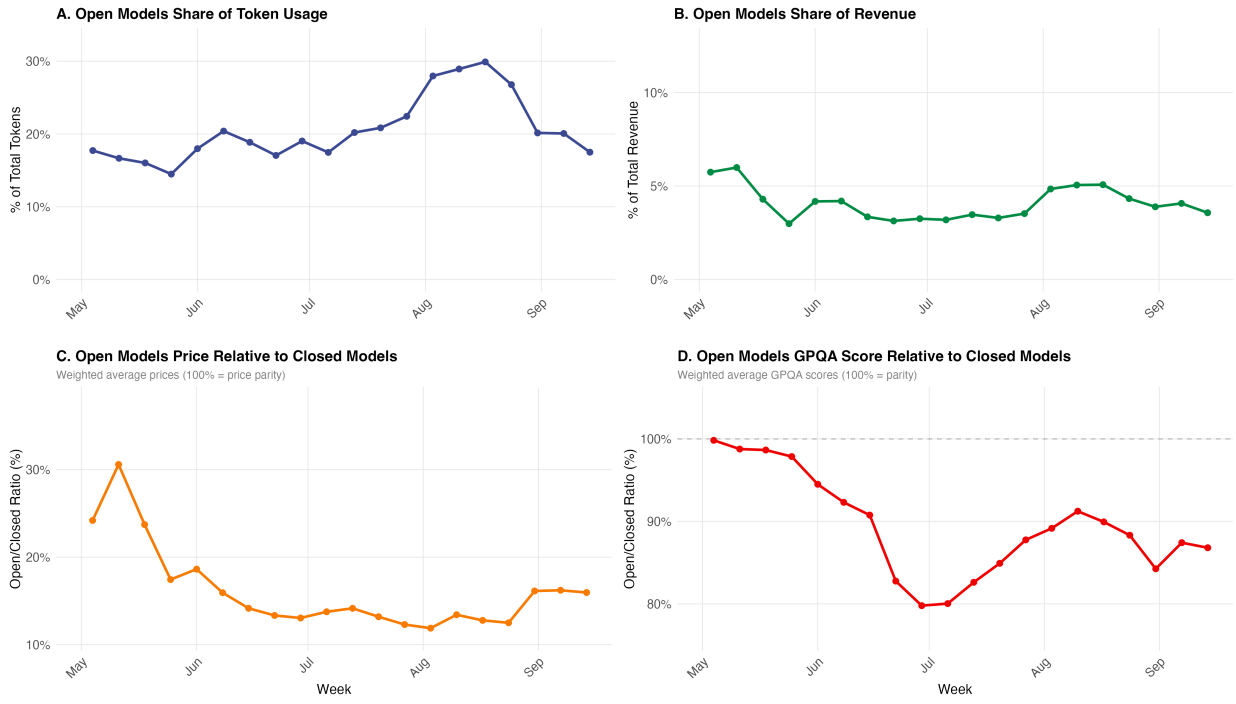

Figure 3: Open Model Market Shares and Relative Performance. Panels A and B show weekly market shares from May to September 2025. Panels C and D show usage-weighted averages of open model metrics relative to closed models (100% = parity).

3.1.2 Open Models Rapidly Match Closed Model Performance

One obvious hypothesis for why closed models command both higher usage and higher prices is that they offer superior performance. To evaluate this hypothesis, we examine the evolution of model capabilities over time and quantify both the performance gap between open and closed models and how quickly open models catch up to closed model performance.

Figure 4 illustrates the evolution of frontier model capabilities on the GPQA benchmark from May 2023 to September 2025. The figure connects models from the same model family, showing the increasing performance of both open and closed model families over this period, as well as the gap between them[21]. Leading closed models (shown in grey) typically introduce new capability levels, which are then matched by open models (shown in color) within months. This pattern is consistent across different model families, with Chinese AI labs like DeepSeek and Qwen, as well as Meta's Llama series and OpenAI's gpt-oss series, demonstrating rapid capability improvements that narrow the gap with closed models. Figure A5 in the appendix shows that this pattern is robust to the choice of benchmark, with similar dynamics observed for MMLU Pro performance.

To quantify these dynamics more precisely, we compute two complementary metrics. First,

[21] To provide context for these performance levels: in our own classroom experience, students typically answer 0 to 2 questions correctly out of 10 on GPQA benchmarks, illustrating both the difficulty of these graduate-level questions and the astonishing capability that these models have achieved over this relatively short time period.

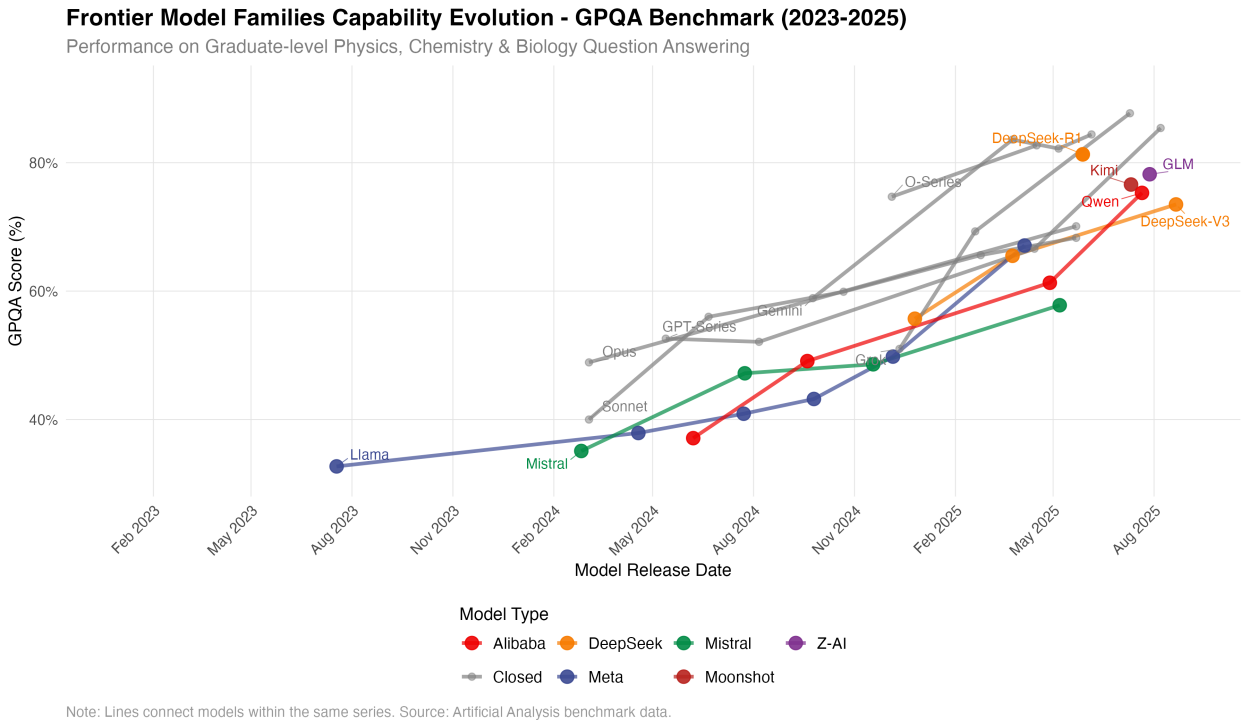

Figure 4: Frontier Model Families GPQA Performance Evolution (2023-2025). GPQA scores for leading model families over time, showing the evolution of both closed (grey) and open (colored) models. Lines connect models within the same family, with each point representing a model release. The figure illustrates both the increasing capabilities over time and the capability gap between open and closed models.

we measure *catch-up time*[22]: for each week, we identify the best-performing closed model and calculate how many weeks elapse before an open model achieves the same or better score. Second, we measure the *capability gap*: the percentage difference in performance between the best open and closed models available at each point in time. For GPQA and LiveCodeBench (which have maximum scores of 100%), the gap is calculated relative to perfect performance; for LM Arena scores (which have no natural maximum), the gap is normalized by the all-time best score observed in our data.

Figure 5 presents these metrics across three leading benchmarks: GPQA (graduate-level science questions), LiveCodeBench (coding challenges), and LM Arena - Text (conversational abilities). The top row shows catch-up time, while the bottom row shows the capability gap. The horizontal lines in the top panels represent average catch-up times across six-month periods, while the horizontal lines in the bottom panels show overall average gaps for each benchmark.

The catch-up time results (top row of Figure 5) reveal that open models are fast followers, with catch-up times that have been decreasing over our analysis period. The sawtooth pattern emerges naturally: after each frontier closed model is released, the catch-up time declines as the date when

[22]Our approach to measuring catch-up time was inspired by the analysis done by Epoch AI (2025a).

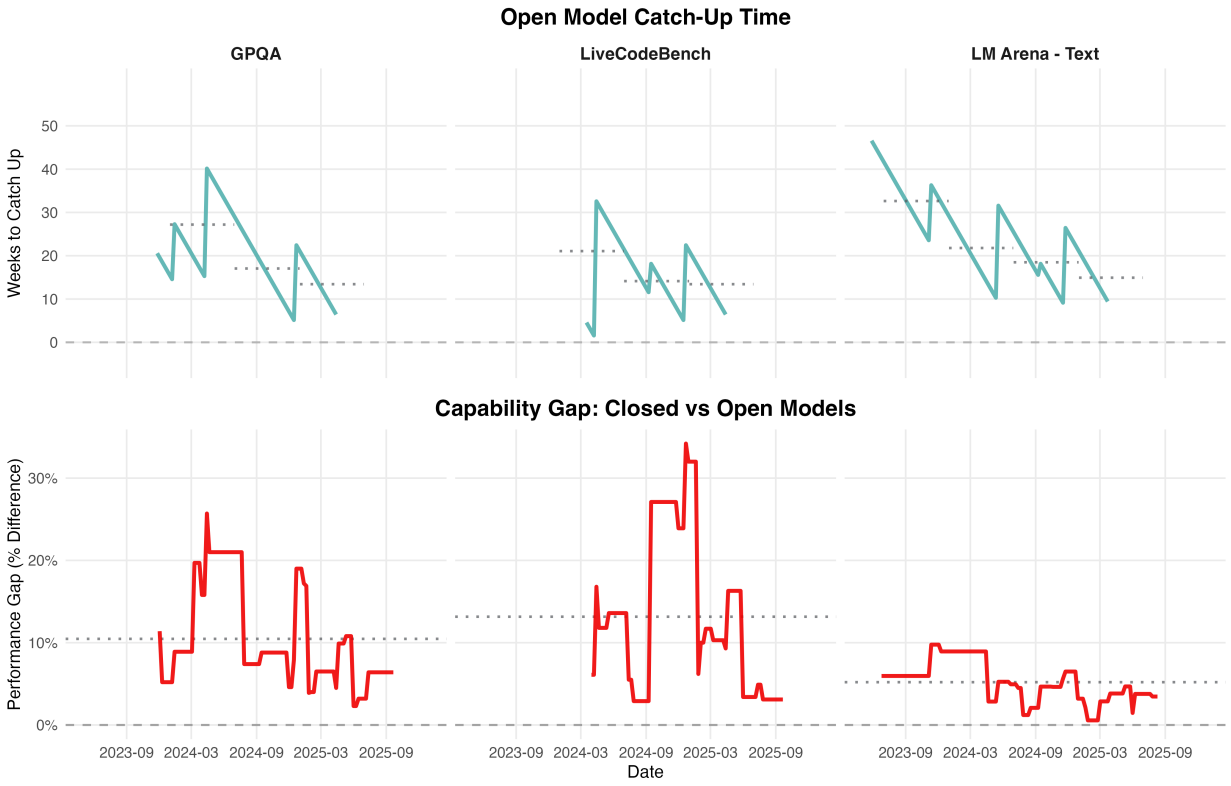

Figure 5: Open Model Catch-Up Time and Capability Gap. The top row shows the time (in weeks) it takes for the leading open model to catch-up to the same performance benchmark as the leading closed model at the time. The bottom row shows the capability gap between the leading open model and closed model at a given time - the percentage on the vertical access shows how much lower the leading open model was than the leading closed model on a given benchmark. Both show the time period from June 2023 to September 2025, with the time series starting based on when the benchmarks were first published. The columns show the performance on three different leading benchmarks used to evaluate LLMs (details discussed in the Methods section).

a frontier open model will surpass it draws closer, then resets when the next frontier closed model appears. For GPQA, the average catch-up time was 27 weeks in the first half of 2024, decreased to 17 weeks in the second half of 2024, and further declined to 13 weeks in the first half of 2025. This pattern of accelerating catch-up is consistent across benchmarks. Open models now routinely match closed model performance within three to six months, and this window continues to shrink. The recent rapid catch-up appears driven by increased investment in open model development from Chinese AI labs including Alibaba (Qwen), DeepSeek, and Z.ai (GLM), as well as continued releases from Meta's Llama series.

The capability gap results (bottom row of Figure 5) reinforce that while closed models maintain a performance edge, the gap is modest and shrinking. For GPQA, the largest performance gap in our sample period was 25.7% (observed in April 2024), but the gap has since narrowed considerably, averaging 10.5% over the full period. This result is robust across MMLU Pro (93.0%), Live-

CodeBench (83.3%), and LMArena (97.25%) over the same analysis period. On the LiveCodeBench coding benchmark, gaps occasionally spike to 30% when new closed models are released, but quickly compress as open models catch up. On the LM Arena - Text benchmark measuring conversational abilities, the gap has remained below 10% since early 2025, with periods where open and closed models achieve near parity.[23] Overall, these results demonstrate that while closed models consistently lead in capabilities, open models deliver performance that is typically 90% or more of closed model performance, with that gap closing rapidly over time.

3.1.3 Open Models Enable Inference Provider Competition that Lowers Prices

While modest performance differences may partially explain a price gap between open and closed models documented above, competition plays a more important role. The fundamental difference in how open and closed models can be deployed creates starkly different competitive dynamics for inference provision. Closed models, by definition, can only be served by their creating companies or licensed partners. In contrast, open models can be deployed by any inference provider with sufficient computational resources—including specialized third-party providers such as Groq, Cerebras, FriendliAI, and Together AI (introduced in the Methods section). This structural difference implies that open models should attract significantly more competing providers than closed models.

Consider the following concrete examples from OpenRouter (Figure A6 provides screenshots of OpenRouter's user interface). OpenAI's GPT-5, a leading closed model, is offered exclusively by OpenAI and its commmercial partner Microsoft Azure[24]. Meanwhile, DeepSeek-V3, a high-performing open model, is available through 16 different inference providers, each competing on price, latency, and service quality. Basic economic theory predicts that as the number of competing providers increases, prices should decline toward marginal cost. For closed models, the monopolistic (or near-monopolistic) market structure allows providers to maintain higher markups. For open models, competitive pressure among multiple independent providers should drive prices down.

Figure 6 demonstrates this theoretical prediction holds (in equilibrium) empirically. The figure plots each model's price per million tokens against the number of inference providers serving that model, with closed models shown as red circles and open models as blue triangles. The pattern is striking: open models cluster at low prices (most below $1 per million tokens, many below $0.10) and high provider counts (up to 19 providers), while closed models occupy the high-price, low-provider region (many above $10 per million tokens, with as few as one provider). The negative correlation between provider count and price is evident for open models, with a correlation of -0.23 between log price and number of providers. This indicates that open models with more providers systematically command lower prices, consistent with competitive dynamics driving prices

[23] Comparing percent differences between LM Arena's benchmarks and the benchmarks from artificial analysis is challenging because of different scales. Artificial analysis benchmarks like GPQA or MMLU Pro tend to be on a zero-to-one scale, whereas LM Arena values are in the thousands.

[24] Even when closed models have multiple providers, these arrangements typically involve the model creator and a small number of cloud partners operating under coordinated pricing. For example, Anthropic's Claude Sonnet 4 is offered by five providers (Anthropic, Amazon Bedrock, Google Vertex US/EU/Global), but all charge identical prices, suggesting the model creator sets the price rather than true market competition.

toward marginal cost. In contrast, closed models exhibit a slightly positive correlation between provider count and price, reflecting that additional providers for closed models are typically second-party licensees (such as cloud partners) that coordinate pricing with the model creator rather than competing independently—making provider count an indicator of market scale rather than competitive pressure.

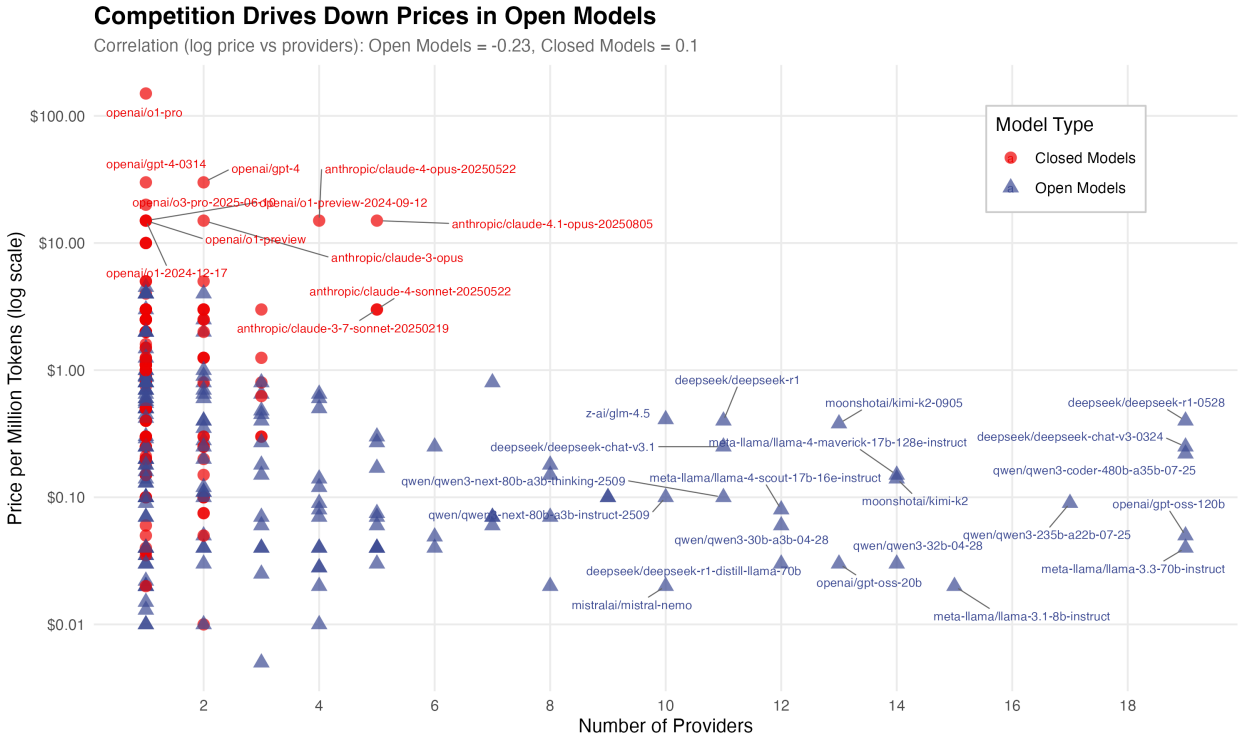

Figure 6: Competition Drives Down Prices in Open Models. This data comes from OpenRouter and shows the number of inference providers for each model in September 2025. The vertical axis shows the price per million tokens (on a log scale) for each model. Red circles represent closed models, blue triangles represent open models. For models with more than one provider, the price shown is the lowest price available on OpenRouter across all providers.

These pricing differences are substantial. Constructing usage-weighted averages across all models in Summer 2025, closed models average $1.86 per million tokens while open models average $0.23 per million tokens—an 87% price reduction. Nearly all open models price below $1.00 per million tokens, with many below $0.10, while closed models frequently exceed $10.00 per million tokens (such as Anthropic's opus model or OpenAI's gpt-4 model), reaching as high as $150.00 (OpenAI's o1-pro model). Beyond driving down open model prices directly through within-model competition, the existence of cheap open alternatives likely constrains closed model pricing through cross-model competitive pressure, although we cannot rule out alternative dynamic explanations like predatory pricing. Without open models providing a low-cost alternative, closed model providers would face less constraint on their ability to increase prices.

Taken together, the results from the previous three subsections establish a puzzle. Closed models dominate the LLM inference economy, capturing approximately 80% of token usage and 96% of revenue despite commanding substantially higher prices. Yet this dominance persists even though open models deliver comparable performance—averaging 90% of closed model capabilities on leading benchmarks—and this performance gap is both modest and shrinking, with open models now catching up to frontier closed models within 3-6 months rather than the 6+ months observed in early 2024. Moreover, the price differential cannot be explained by marginal cost differences alone: open models average 87% lower prices than closed models, a gap driven primarily by competitive market structure rather than fundamental technological advantages. This raises a central question: if open models offer nearly equivalent performance at dramatically lower prices, why do closed models continue to dominate usage? The persistence of this pattern suggests either that users are systematically undervaluing open models relative to their true price-performance characteristics, or that closed models provide benefits beyond raw benchmark performance that justify their price premium. While we cannot directly asnwer this question in this paper, we turn to quantifying its economic significance in the following subsection.

3.2 Open Models Are Underutilized at Economically Significant Levels

To this point, our analysis has focused on describing the state of the LLM inference economy and understanding the quantifiable differences between open and closed models. Now, we turn to quantifying the economic significance of the puzzle established earlier: if open models offer comparable performance at substantially lower prices, why do closed models continue to dominate? This pattern suggests significant underutilization of open models relative to their observable price and performance characteristics, though this need not imply true market inefficiency—rather, it indicates that choice models driven solely by observable metrics are incomplete, with additional unobserved factors influencing selection. We frame our estimation exercise as quantifying the implicit value consumers place on these unobserved factors through three steps: first, regression analysis quantifying the conditional correlation between openness and usage after accounting for price and performance; second, visualizing patterns of observable domination; and third, simulating consumer savings from switching observably dominated closed models to superior open alternatives.

3.2.1 Observational Regression Analysis of Model Usage

We begin by formally estimating the relationship between model openness and usage, controlling for price and benchmark performance. Our baseline specification regresses log token usage on log price, an open source indicator, and standardized benchmark scores:

$$\log(\text{Tokens}_m) = \beta_P \log(\text{Price}_m) + \beta_B \mathbf{Benchmarks}_m + \beta_O \text{Open}_m + \varepsilon_m \tag{1}$$

where Tokens_m represents total token usage for model m in July 2025, Price_m is the price per million tokens, $\mathbf{Benchmarks}_m$ is a vector of standardized benchmark scores, and Open_m is an

indicator for open source models (centered at 0.5 for open models and -0.5 for closed models to facilitate interpretation).

We emphasize that these regressions are descriptive and do not identify causal effects. Model usage is endogenous to numerous unobserved factors including brand reputation, developer ecosystem, API reliability, safety features, and user preferences that are correlated with both openness and usage. Our goal is not to estimate causal effects but rather to quantify the negative conditional correlation between openness and usage after accounting for the most obvious observable characteristics—price and capability as measured by standard benchmarks[25].

A further concern is whether missing benchmark data for low-performing models affects these results. In A.2, we explore how the selection of models into the benchmarked sample might bias our estimates, arguing that this missingness likely causes us to *underestimate* rather than overestimate the magnitude of the negative correlation between openness and utilization.

In preparing the data for regression analysis, we standardize all benchmark scores by subtracting their sample mean and dividing by their standard deviation. This ensures that coefficients are comparable across benchmarks with different scales. The open source indicator is centered rather than using a standard 0/1 dummy variable, which allows the intercept to represent the average across both model types. We focus our analysis on models with positive usage and non-missing price data in July 2025, yielding sample sizes ranging from 82 to 264 models depending on benchmark data availability.

Figure 7 presents the estimated coefficients on the open source indicator across seven different model specifications. The specifications vary in three dimensions: (1) the set of benchmarks included (baseline with no benchmarks, Artificial Analysis benchmarks only, LMArena benchmarks only, or all available benchmarks), (2) whether interaction terms between openness and benchmarks are included, and (3) whether the log price is included as a control. Across all specifications, we consistently find a statistically significant negative coefficient on the open source indicator for usage regressions, ranging from approximately -1.0 to -2.1. Translating these coefficients to percentage effects, this implies that open models receive approximately 63% to 88% lower usage than comparable closed models ($e^{-1.0} - 1 = -0.63$ to $e^{-2.1} - 1 = -0.88$). This indicates that open models receive substantially lower usage than closed models, even after controlling for their price advantage and benchmark performance. The pattern is robust across different benchmark sets and model specifications. Full regression tables showing coefficients for all variables are presented in Table A3 and Table A4 in the appendix. The consistent negative coefficient on openness suggests that factors beyond raw performance metrics drive model selection.

3.2.2 Many Closed Models Have Better and Cheaper Open Alternatives

The regression results motivate a closer examination of the raw data to understand what drives the underutilization of open models. Figure 8 plots models in price-performance space for two key benchmarks (GPQA and LiveCodeBench) during summer 2025, with bubble sizes proportional to

[25]Developing better identification strategies for model demand remains an important area for future work.

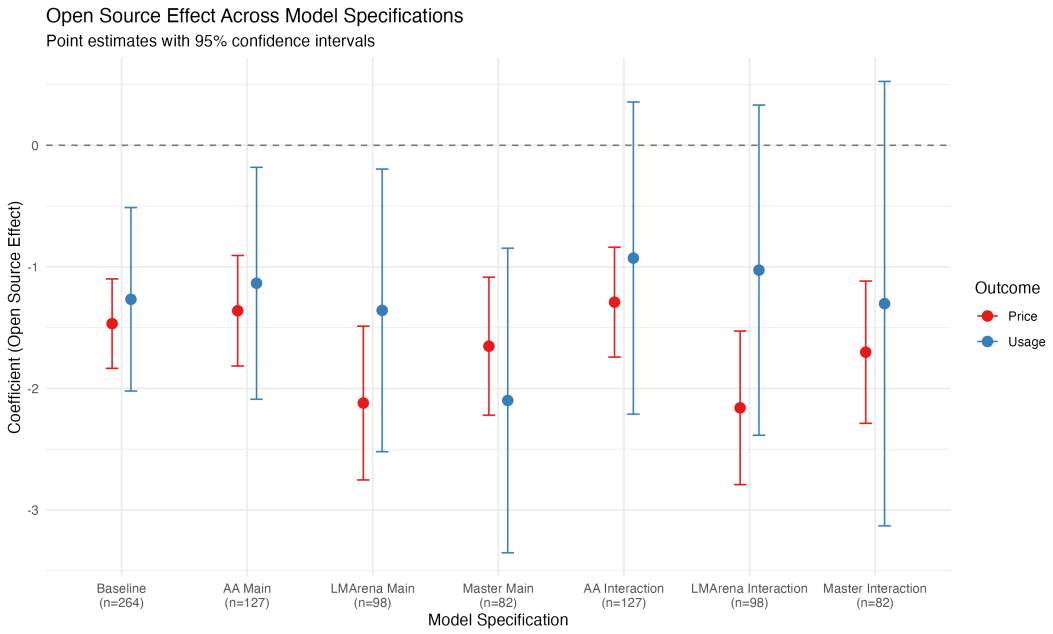

Figure 7: Open source effect on price and usage across model specifications. Point estimates with 95% confidence intervals. All specifications show a negative effect of openness on usage, indicating that open models are systematically underutilized relative to their price and performance characteristics. Models with interacted terms included lose statistical significance but are directionally consistent and maintain a large point-estimate magnitude. Sample sizes vary from n=82 to n=264 depending on benchmark data availability.

token usage. The fact that many closed models have superior open alternatives is unsurpising, given the pace of development of technological capabilities in this market for both open and closed models. However, it is surprising that many highly-used closed models are observably dominated by open alternatives that offer both superior performance and lower prices. The full visualization across all benchmarks (MMLU Pro, GPQA, LiveCodeBench, and LM Arena scores) is presented in Figure A7 in the appendix, showing that this pattern is robust across different measures of capability.

For example, the open source model Qwen 235B achieves higher scores than GPT-4.1-2025-04-14 on both GPQA (0.75 vs 0.67) and LiveCodeBench (0.52 vs 0.46) while costing approximately one-tenth the price ($0.18 vs $2.00 per million tokens). Yet GPT-4.1 receives orders of magnitude more usage. Similar patterns appear across multiple model pairs. Much of this observable domination comes from a small set of Chinese open source models released in late 2024 and early 2025: DeepSeek R1 and DeepSeek V3, Qwen 3 235B, Kimi K2, GPT OSS 120B, and GLM-4.5. Conversely, much of the dominated usage is concentrated in popular closed models from major Western providers: Claude Sonnet 4 and Claude Sonnet 3.7, Gemini 2.5 Flash and Gemini 2.0 Flash, Grok Code Fast 1, and GPT-4o-mini.

These patterns of observable domination suggest substantial unrealized consumer savings, if we believe choice is determined by price and benchmark score. If users were to switch from domi-

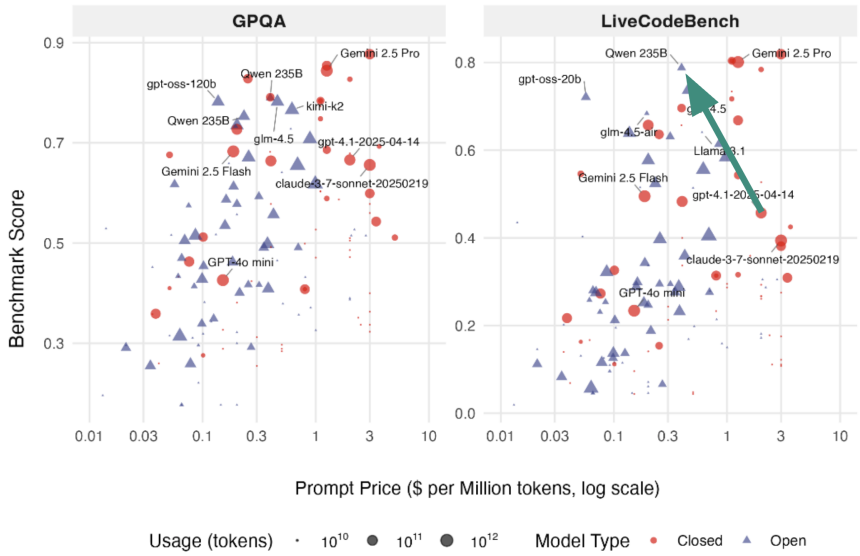

Figure 8: Price-performance frontier for open versus closed models, summer 2025. Models in the upper-left region (high performance, low price) dominate those in the lower-right. Notable examples include Qwen 235B dominating GPT-4.1 on both dimensions, yet GPT-4.1 receiving substantially higher usage (indicated by bubble size). The green arrow illustrates the improvement from making this switch - benchmark perforance goes up and cost goes down.

nated closed models to dominating open alternatives—receiving both better performance and lower prices—they could realize significant savings while also improving functionality. This motivates our final analysis: quantifying the magnitude of this unrealized value through a counterfactual simulation.

3.2.3 Estimating and Extrapolating the Unrealized Value of Open Models

To quantify the economic significance of open model underutilization, we conduct a counterfactual simulation exercise. The core idea is simple: for each week and each closed model used during our observation period (June-August 2025), we identify whether a superior open alternative exists during that week—one that both outperforms the closed model on capability benchmarks and offers a lower price. When such a dominating open model exists, we simulate switching all usage from the dominated closed model to its superior open alternative. This constitutes an improvement where users who switch experience both better performance and lower costs, while users of closed models without superior open alternatives remain unaffected.

We implement three matching strategies to identify appropriate open model matches if there are multiple possible choices, varying in how they balance the price-performance tradeoff: (1) *Best Performance* selects the highest-performing open model that exceeds the closed model's benchmark score, (2) *Lowest Price* selects the cheapest open model that still outperforms the closed model, and (3) *Best Value* selects the open model optimizing the price-performance ratio while maintaining superior performance. For each strategy, we compute the analysis separately using four leading

benchmarks (GPQA, LM Arena, LiveCodeBench, and MMLU Pro), then aggregate results by computing usage-weighted averages across benchmarks. For each dominated closed model, we calculate the imputed cost savings by multiplying its token usage by the price difference between the closed model and its open alternative. Summing across all dominated models yields the total unrealized value—the consumer savings that could be realized through optimal model selection.

The simulation results are presented in Table A1 in the appendix. Across all matching strategies and benchmarks, we consistently find that switching from dominated closed models to superior open alternatives would greatly reduce average prices (averaging 70.7% across matching methods and benchmarks) while simultaneously improving average benchmark performance greatly (averaging 14.3%). This pattern—large cost savings coupled with capability improvements—validates that we are identifying substantial improvements rather than incremental ones. The consistency of these results across different benchmarks and matching strategies suggests our findings are robust to methodological choices.

The key output of this simulation is the magnitude of revenue savings—or equivalently, unrealized consumer savings. Figure 9 presents these results across all matching strategies and benchmarks for our 14-week observation period. The figure shows both actual revenue (darker bars) and imputed revenue under optimal switching (lighter bars), with percentage reductions and absolute savings labeled. Revenue savings range from 57.4% to 80.0% across method-benchmark combinations, with a median of 70.6%. For our 14-week summer 2025 observation period, these savings range from \$27.8M to \$37.8M, with a median of \$32.2M. Extrapolating to a full year yields unrealized value estimates of \$104M to \$146M annually on OpenRouter alone, representing approximately 57-80% of the platform's total annual revenue. These ranges—57.4%, 70.6%, and 80.0%—form the parameters we use to extrapolate to the broader market, providing conservative, median, and aggressive estimates respectively of open model underutilization.

To assess the broader economic significance of these findings, we must extrapolate from OpenRouter to the global LLM inference market. This extrapolation requires estimating what fraction of total market activity OpenRouter represents—a challenging task given the limited public data on this rapidly evolving market. We employ three independent approaches using different reference values: (1) Menlo Ventures' first half of 2025 market size estimate based on surveys of technical leaders at 150 companies (Tully et al., 2025) and an extrapolation to all of 2025 via an exponential growth model (detailed in A.3), (2) publicly disclosed token volumes from Google (Hassabis, 2025), (3) estimated OpenAI API revenue derived from Epoch AI's total revenue estimates (Epoch AI, 2025b) combined with industry-reported API revenue fractions (Future Search, 2025; Khaan, 2025). For (2) and (3), we compute the analogous quantities observed on OpenRouter (i.e., Google tokens or OpenAI revenue) in order to infer the OR fraction of market, and then use that fraction to scale up observed OR revenue to the total market. We focus on the Menlo Ventures estimate, which we consider most reliable for several reasons. First, it derives from actual spending data collected from technology leaders rather than partial metrics like token counts or single-company revenue. Second, it focuses specifically on the API inference market rather than requiring assumptions about

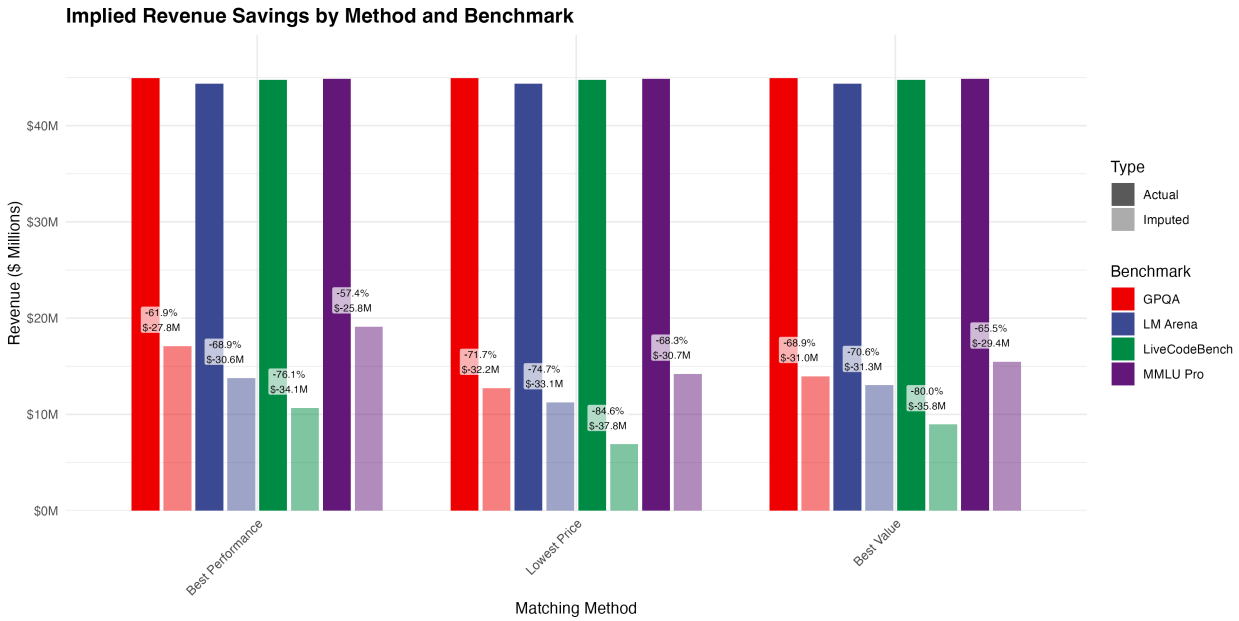

Figure 9: Implied Revenue Savings by Matching Method and Benchmark. Darker bars show actual revenue during summer 2025; lighter bars show imputed revenue if all dominated closed models switched to superior open alternatives. Labels indicate percentage reduction and absolute savings in millions. Revenue savings range from 57.4% to 80.0% across methods, with a median of 70.6%, representing $27.8M to $37.8M in unrealized consumer savings over the 14-week observation period.

what fraction of broader metrics represent API usage. However, this approach is well-justified by the rapid market expansion and yields a conservative $35.1B estimate for 2025, lower than the value provided by the other methods. The detailed reference values, OpenRouter fractions, and calculation methodology are documented in Table A2 in the appendix. Notably, this table shows that OpenRouter comprises between $0.3\% - 1.06\%$ of the various reference values, giving us a sense of the total market share.

These three approaches yield different market size estimates for 2025: $42.7B (Google tokens), $60.4B (OpenAI API revenue), and $35.1B (Menlo Ventures). We emphasize that these are order-of-magnitude estimates intended to contextualize our findings rather than precise market measurements. Our assumption that OpenRouter usage patterns are representative of the broader market is almost certainly imperfect, as OpenRouter may attract more technically sophisticated users, and in particular those that value interoperability more compared overall market (see 2.2.1) — in this case, it would imply that our results understate the degree of open model underutilization.

With these market size estimates in hand, we can extrapolate our OpenRouter findings by assuming representativeness. While this is a very strong assumption (with nuances discussed in 2.2.1), it nevertheless allows us to generate a ball-park estimate of the unrealized value of open models in the broader AI economy. Table 1 presents the complete set of extrapolation results. The top two rows illustrate the realized underutilization on OpenRouter (without any out-of-

sample extrapolation assumptions). The bottom three rows combine our three market size estimates with our three underutilization rates (57.4%, 70.6%, and 80.0% from the Best Performance, Best Value, and Lowest Price matching strategies). Using our preferred Menlo Ventures market size estimate of $35.1B, the unrealized value of open models ranges from $20.1B to $28.1B annually, depending on the assumed underutilization rate, with a median estimate of $24.8B (our prefererred estimate). The other extrapolation methods yield higher estimates: $24.5B to $34.2B using Google token data, and $34.7B to $48.3B using OpenAI API revenue estimates. Across all estimation approaches, the central tendency suggests that optimal substitution to open models could save the AI industry approximately $25B annually—representing roughly 70% of total market spending. These magnitudes underscore that the underutilization of open models represents a first-order economic phenomenon rather than a marginal inefficiency.

Table 1: Estimated Savings from Open Models Extrapolated to Total Market

			Estimated 2025	Underutilization Value		
Concept	Time	Method	Market Value	57.4%	70.6%	80.0%
Open Router Market	Summer 2025	Observed	$49M	$28M	$34M	$39M
Open Router Market	2025	Extrapolated from Observed	$182M	$104M	$128M	$146M
Total Inference Market	2025	Extrapolated from Menlo Ventures	$35.1B	$20.1B	$24.8B	$28.1B
Total Inference Market	2025	Extrapolated from OpenAI API Revenue Estimates	$60.4B	$34.7B	$42.6B	$48.3B
Total Inference Market	2025	Extrapolated from Google Token Estimates	$42.7B	$24.5B	$30.2B	$34.2B

Note: This table shows market size extrapolations and potential savings from switching to observably superior open source models. Open Router Market rows show observed (14-week period June-August 2025) and annualized revenue. Total Inference Market rows show market size estimates from three different extrapolation methods: Menlo Ventures H1 2025 estimate ($8.6B, annualized to $35.1B), OpenAI API revenue (Epoch AI total revenue data: $10B ARR in May, $12B ARR in July, with 20% API fraction), and Google token counts (Demis tweet). Underutilization Value columns show potential cost savings at different underutilization rates (57.4%, 70.6%, 80.0%) corresponding to Best Performance, Best Value, and Lowest Price matching strategies from the unrealized value analysis.

To provide additional context on the value open models currently deliver, we also conducted a complementary analysis examining the reverse scenario: what would users of open models pay if open models did not exist and they had to switch to the cheapest closed alternative with equivalent or better performance? This calculation mirrors prior research valuing unpriced open source software by estimating the cost of proprietary substitutes (e.g., Hoffmann et al., 2024). We find that the absence of open models would increase spending by open model users by 1.0-3.5% of total OpenRouter revenue (depending on benchmark and matching method), constituting $350MM-$1.23B actual value saved across the total inference market. This is more than an order of magnitude smaller than unrealized value from dominated closed model usage. This asymmetry occurs because: (1) open models are used substantially less than closed models, so simulated changes to the behavior has less of relative impact on the total market (2) many open models have cheaper closed alternatives than the most popular closed models on the market — presumably distilled versions of the highest performance closed models (like the Gemini Flash Lite series). In other words, some of the proposed inefficiency in not choosing open models could also be addressed by simply switching some closed model usage to better and cheaper (and presumably newer) closed models.

4 Discussion

In aggregate, perhaps the most striking finding from our analysis is not the value that open models currently create, but rather the far larger value they could create if users made (potentially) more efficient choices. The substantial underutilization we document—with users continuing to rely on closed models even when observably superior open alternatives exist—suggests that closed model providers maintain meaningful market power despite the availability of similarly capable open alternatives at dramatically lower prices. While our analysis quantifies this underutilization, the precise mechanisms through which closed models retain their dominant position remain an open question for future research. Several potential channels merit consideration, which we organize in Table 2.

Table 2: Potential Explanations of Open Model Underutilization

Driver	**Description**
Unobserved Consumer Preferences	
Benchmark mismeasurement	Standardized benchmarks may not capture quality dimensions that matter in production (e.g., safety, alignment, style, hallucinations), and closed models may be superior on those dimensions
Switching costs	Migration costs from workflows adapted to specific model behaviors, prompt optimization, or proprietary features
Liability backstop	Established providers offer legal recourse and accountability that may be absent with smaller inference providers
System diversification	Router-based architectures benefit from maintaining access to multiple model families and providers
Security concerns	Data exfiltration risks from foreign-trained models or untrusted inference providers
Experimentation dynamics	Open models used for development or last-mile optimization rather than initial production workloads
Information Frictions	
Inattentiveness to model innovation	Lack of awareness of cheaper and higher performance alternatives, possibly due to pace of change, leads to use of outdated models
Brand and organizational risk aversion	"No one ever got fired for buying IBM/Microsoft/OpenAI"—institutional conservatism favors established vendors as has been the case for decades
Misconceptions about open models	Beliefs that using open models exposes proprietary data to competitors or the public
Misconceptions about foreign models	Confusion between foreign models and foreign inference providers regarding data privacy implications

These mechanisms broadly fall into two categories: unobserved consumer preference, unseen by the researcher but known by the consumer, and information frictions that could potentially be mitigated through policy or education. Among the preference-based explanations, benchmark mismeasurement appears particularly salient—standardized tests may fail to capture quality dimen-

sions like safety or reliability that may matter in production settings. Switching costs also merit attention, as systems and workflows become adapted to the idiosyncratic behaviors of specific models through extreme prompt optimization, making migration costly despite potential savings. While such factors may be difficult to mitigate with policy and suggest valid economic reasons driving our alleged underutilization, our contribution here is to point out that the economic value of these factors is surprisingly large and likely drives the market power seen by closed model providers (as opposed to pure capabilities leadership). Information frictions may be equally important, and constitute potential inefficiencies. We emphasize that such switching costs are conceptually distinct from simple inattentiveness to model innovation, which may explain use of older systems but are potentially ameliorated through information interventions. Further, in anecdotal conversations with organizations adopting AI, we have encountered persistent misconceptions—from beliefs that using open models will expose proprietary data to competitors, to concerns that queries to foreign-trained models might be intercepted by foreign operatives even if they are hosted locally. These misunderstandings echo the early days of open source software, when the common refrain "no one ever got fired for buying IBM" reflected deep-seated organizational conservatism. That principle evolved into "no one ever got fired for buying Microsoft" during the operating system battles of the 1990s and 2000s, and today one might readily hear "no one ever got fired for buying OpenAI/Anthropic/Google." Understanding the relative importance of these mechanisms represents a crucial direction for future work, as it would illuminate both the durability of current market structure and the potential for efficiency gains through better information or reduced frictions.

Perhaps equally surprising as the underutilization of open models is the sheer volume and capability of these models in the first place. Given the enormous fixed costs associated with training frontier AI models—investments measured in hundreds of millions or even billions of dollars—the existence of highly capable open models represents a significant puzzle for economic theory. Understanding the sources and motivations behind open model creation emerges as a critical area for future research. As illustrated in Figure A3, all of the high-usage models in our dataset, whether open or closed, originate from companies rather than academic institutions, nonprofit organizations, or individual developers. This corporate dominance of model production challenges narratives that frame open models as primarily community-driven efforts analogous to traditional open source software. Interestingly, many companies pursue dual strategies, releasing both open and closed models as part of their broader competitive approach. Meta's Llama family, DeepSeek's various releases, Alibaba's Qwen models, and Mistral AI's portfolio all exemplify this hybrid strategy. However, the current market leaders in closed model revenue—OpenAI, Anthropic, Google, and X.ai —have historically focused almost exclusively on closed model development, with notable exceptions being Google's Gemma family and OpenAI's more recent GPT-OSS releases. Finally, the most powerful open models released in 2025 come predominantly from Chinese companies like Alibaba (Qwen), DeepSeek, Moonshot AI (Kimi), and Z.ai (GLM). These observations raise fundamental questions: What enables companies to justify the investments required to train open models when the direct revenue capture mechanisms appear limited? What knowledge-spillover

mechanisms exist that allow open models to catch-up to closed models at lower costs than the original development? What competitive advantages do open model releases confer that offset their substantial costs? And how do the incentives and capabilities for open model production vary across different market positions and geographic regions? Answering these questions will be essential for predicting the long-term sustainability and evolution of the open model ecosystem.

In total, our findings reveal that open models play a fundamentally important, but latent, role in the AI economy. While they account for less than 5% of current inference revenue, our analysis suggests they could deliver $20-$48 billion in consumer savings annually if adopted wherever they match or exceed closed alternatives on observable metrics. This gap—between minimal realized value and substantial potential value—defines their position in today's market: technically capable, economically significant in potential, yet largely untapped. Understanding what drives this pattern is critical for predicting AI market structure, as the answer will determine whether open models remain peripheral or move toward the center of how organizations deploy artificial intelligence. If history is any guide, just as organizations eventually came to understand and embrace the value created by open source software, they will likely come to understand the value created by open models, shifting unrealized value to realized value and unlocking even greater economic impact than we see today.

References

Daron Acemoglu and Simon Johnson. Big tech is bad. big a.i. will be worse. 2023. URL https://www.nytimes.com/2023/06/09/opinion/ai-big-tech-microsoft-google-duopoly.html. Accessed: 2025-11-17.

Artificial Analysis. Intelligence benchmarking methodology, 2025. URL https://artificialanalysis.ai/methodology/intelligence-benchmarking.

Pierre Azoulay, Joshua L. Krieger, and Abhishek Nagaraj. *Old Moats for New Models: Openness, Control, and Competition in Generative Artificial Intelligence*. University of Chicago Press, 2025. URL https://www.nber.org/books-and-chapters/entrepreneurship-and-innovation-policy-and-economy-volume-4/old-moats-new-models-openness-control-and-competition-generative-artificial-intelligence.

William P. Barr. Big tech's budding ai monopoly. 2024. URL https://www.wsj.com/articles/big-techs-budding-ai-monopoly-40280c15. Accessed: 2025-11-17.

Knut Blind, Mirko Peti, Matthias Böhm, Ariel Katz, Sachiko Muto, Matthias Stürmer, Nils Sachse, and Torben Schubert. The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy. Technical report, Publications Office of the European Union, Luxembourg, 2021. URL https://digital-strategy.ec.europa.eu/en/library/study-about-impact-open-source-software-and-hardware-technological-independence-competitiver

Cerebras Systems. Cerebras: Wafer-scale AI inference and training, 2024. URL https://www.cerebras.ai/.

Harrison Chase and LangChain Contributors. LangChain: Building applications with LLMs through composability, 2024. URL https://github.com/langchain-ai/langchain.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024. URL https://arxiv.org/abs/2403.04132.

Avinash Collis and Erik Brynjolfsson. Ai's overlooked $97 billion contribution to the economy. *The Wall Street Journal*, 2025. URL https://www.wsj.com/opinion/ais-overlooked-97-billion-contribution-to-the-economy-users-service-da6e8f55.

Mert Demirer, Andrey Fradkin, Nadav Tadelis, and Sida Peng. The emerging market for intelligence: The supply, demand, and usage of llms. 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. doi: 10.1126/science.adj0998.

Epoch AI. Open models: A new frontier in AI development. Technical report, Epoch AI, January 2025a. URL https://epoch.ai/blog/open-models-report.

Epoch AI. AI companies database, 2025b. URL https://epoch.ai/data/ai-companies.

Fal AI. fal.ai: Generative media platform for developers, 2024. URL https://fal.ai/.

Andrey Fradkin. Demand for LLMs: Descriptive evidence on substitution, market expansion, and multihoming. *arXiv preprint arXiv:2504.15440*, April 2025. URL https://arxiv.org/abs/2504.15440.

FriendliAI. FriendliAI: The generative AI infrastructure company, 2024. URL https://friendli.ai.

Future Search. Revenue model analysis: OpenAI and Anthropic. Technical report, Future Search, 2025. URL https://app.futuresearch.ai/reports/3Li1.

Shane Greenstein and Frank Nagle. Digital dark matter and the economic contribution of Apache. 43(4):623–631, 2014. doi: 10.1016/j.respol.2014.01.003.

Groq, Inc. Groq: Fast, low-cost AI inference, 2024. URL https://groq.com/.

Mahyar Habibi. Open sourcing GPTs: Economics of open sourcing advanced AI models. *arXiv preprint arXiv:2501.11581*, January 2025. URL https://arxiv.org/abs/2501.11581.

Demis Hassabis. Tweet about Google token processing volumes. Twitter/X, January 2025. URL https://x.com/demishassabis/status/1948579654790774931.

Manuel Hoffmann, Frank Nagle, and Yanuo Zhou. The value of open source software. Strategy Unit Working Paper 24-038, Harvard Business School, 2024. URL https://www.hbs.edu/faculty/Pages/item.aspx?num=65230.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. URL https://arxiv.org/abs/2403.07974.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770.

Benjamin Jones. Artificial intelligence in research and development. Working Paper c15299, National Bureau of Economic Research, 2025.

Paul Kedrosky. Honey, AI capex is eating the economy. Paul Kedrosky Newsletter, July 2025. URL https://paulkedrosky.com/honey-ai-capex-ate-the-economy/.

Furqan Khaan. How OpenAI and Anthropic are cashing in on AI: A look at their revenue models. Medium, 2025. URL https://medium.com/@furqankhaan/how-openai-and-anthropic-are-cashing-in-on-ai-a-look-at-their-revenue-models-d9d9ae79dd28.

Gizem Korkmaz, J Bayoán Santiago Calderón, Brandon L Kramer, Ledia Guci, and Carol A Robbins. From github to gdp: A framework for measuring open source software innovation. *Research Policy*, 53(3):104954, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP'23)*. ACM, 2023. URL https://github.com/vllm-project/vllm.

Matthew Leisten. Open(?) AI. *Available at SSRN 5044391*, January 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5044391.

Dan Milmo. Microsoft, OpenAI and Nvidia investigated over possible breach of antitrust laws. *The Guardian*, June 2024. URL https://www.theguardian.com/business/article/2024/jun/06/microsoft-openai-and-nvidia-investigated-over-possible-breach-of-antitrust-laws.

Frank Nagle. Open source software and firm productivity. *Management Science*, 65(3):1191–1215, 2019. doi: 10.1287/mnsc.2017.2977.

Nomic AI. GPT4All: Run local LLMs on any device, 2024. URL https://github.com/nomic-ai/gpt4all.

Ollama Contributors. Ollama: Get up and running with large language models locally, 2024. URL https://github.com/ollama/ollama.

ONNX Contributors. ONNX: Open neural network exchange, 2024. URL https://onnx.ai/.

Open Source Initiative. Open source AI definition, 2024. URL https://opensource.org/ai/open-source-ai-definition.

Open WebUI Contributors. Open WebUI: User-friendly AI interface for local LLM deployment, 2024. URL https://github.com/open-webui/open-webui.

OpenRouter. OpenRouter raises $40 million to scale up multi-model inference for enterprise. Press Release, June 2025. URL https://www.globenewswire.com/news-release/2025/06/25/3105125/0/en/OpenRouter-raises-40-million-to-scale-up-multi-model-inference-for-enterprise.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019. URL https://arxiv.org/abs/1912.01703.

Dylan Patel and Afzal Ahmad. Google "We have no moat, and neither does OpenAI". SemiAnalysis Newsletter, May 2023. URL https://newsletter.semianalysis.com/p/google-we-have-no-moat-and-neither.

Asad Ramzanali. We need to break up big AI before it breaks us. *TIME*, October 2025. URL https://time.com/7322418/chat-gpt-open-ai-nvidia-tech-monopoly/.

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 Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2024. URL https://arxiv.org/abs/2311.12022.

Toran Bruce Richards and Significant Gravitas Contributors. AutoGPT: An experimental open-source application showcasing GPT-4 autonomy, March 2023. URL https://github.com/Significant-Gravitas/AutoGPT.

Together AI. Together AI: The AI acceleration cloud, 2024. URL https://www.together.ai/.

Tim Tully, Joff Redfern, Deedy Das, and Derek Xiao. 2025 mid-year llm market update: Foundation model landscape + economics. *Menlo Ventures*, 2025. URL https://menlovc.com/perspective/2025-mid-year-llm-market-update/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://arxiv.org/abs/2406.01574.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://arxiv.org/abs/1910.03771.

Fasheng Xu, Xiaoyu Wang, Wei Chen, and Karen Xie. The economics of AI foundation models: Openness, competition, and governance. *arXiv preprint arXiv:2510.15200*, October 2024. URL https://arxiv.org/abs/2510.15200.

Daniel Yue and Frank Nagle. Igniting innovation: Evidence from pytorch on technology control in open collaboration. In *Academy of Management Proceedings*, volume 2025, page 14679. Academy of Management Valhalla, NY 10595, 2025.

A Appendix

A.1 Model Matching and Crosswalk Methodology

A central methodological challenge in our analysis is linking the same underlying AI model across different platforms, each of which employs inconsistent naming conventions. This section provides comprehensive technical details on our crosswalk methodology for matching models between OpenRouter, Artificial Analysis, LM Arena, and Hugging Face.

A.1.1 The Model Naming Problem

The same AI model can appear with substantially different identifiers across platforms. Consider Meta's Llama 3.1 70B Instruct model as an illustrative example. This single model appears as:

- `meta-llama/llama-3.1-70b-instruct` on Hugging Face
- `meta-llama-3.1-70b-instruct` on OpenRouter

- `llama-3.1-70b-instruct` on Artificial Analysis
- `Llama 3.1 70B` on LM Arena

Beyond simple naming variations, several additional complications arise. Models exist in multiple versions distinguished by training checkpoints, with names like `gpt-4-1106-preview` versus `gpt-4-0613` referring to different temporal versions of GPT-4, where these versions may have an undated 'permaslug' that updates with the latest model version. Quantization variants introduce further complexity, with the same base model appearing in 4-bit, 8-bit, and full-precision versions. Instruction-tuning variants multiply model names further, as base models spawn instruct-tuned, chat-tuned, and specialized fine-tuned derivatives. Some platforms include provider-specific identifiers in model names, while others use canonical names. Deprecated models may be renamed or aliased to newer versions.

A.1.2 Multi-Stage Matching Algorithm

To systematically address these challenges, we implement a five-stage matching algorithm that progresses from highest to lowest confidence matches:

Stage 1: Hardcoded Mapping Application. We maintain a manually curated crosswalk file containing verified mappings for models that cannot be reliably matched algorithmically. These hardcoded mappings address several categories of difficult cases. Models with completely different names across platforms (such as `GPT-4 Turbo` versus `gpt-4-1106-preview`) require explicit mapping. Multiple model versions with similar names necessitate careful disambiguation to avoid false matches. Deprecated models that have been renamed or aliased must be tracked through their naming history. Platform-specific quantization or deployment variants that do not correspond to distinct underlying models need consolidation. We developed this hardcoded mapping iteratively by reviewing initial algorithmic matches, identifying errors, and adding manual corrections. Each hardcoded mapping includes documentation of the rationale and evidence supporting the match.

Stage 2: Exact Model Name Matching. For models not resolved by hardcoded mappings, we attempt exact string matching after standardization. We convert all model names to lowercase to eliminate casing differences, normalize whitespace by replacing multiple spaces with single spaces and removing leading/trailing whitespace, and remove common prefixes like `models/` or `openai/` that indicate platform structure rather than model identity. After standardization, we compare model names for exact string equality. This stage successfully matches models with identical naming conventions across platforms.

Stage 3: Slug-Based Matching. Many platforms provide structured identifiers (slugs) that encode model information systematically. Hugging Face uses an `organization/model` format, while OpenRouter often includes similar structured identifiers. For platforms providing slug information, we extract the slug components and attempt matching on the model component after removing organization prefixes. We also handle cases where slugs use different separators (hyphens

versus underscores) by normalizing to a consistent format. This stage captures models where the underlying slug structure is preserved despite surface-level naming differences.

Stage 4: Substring Containment. For models still unmatched, we check whether one model name is a substring of another. We perform bidirectional checks, testing whether name A contains name B and vice versa. This asymmetric approach handles cases where one platform uses an abbreviated form (such as `llama-3.1-70b`) while another uses a fully qualified form (such as `meta-llama-3.1-70b-instruct`). To avoid false positives from short common substrings, we impose a minimum substring length threshold of 10 characters. We also manually review all substring matches to verify they represent the same underlying model rather than different variants.

Stage 5: Levenshtein Distance Similarity. As a final stage for remaining unmatched models, we compute character-level edit distances between candidate model name pairs. The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) needed to transform one string into another. We normalize this distance by the length of the longer string to obtain a similarity score between 0 and 1. We consider candidate matches with similarity scores above 0.85, which captures models differing by minor typos, formatting variations, or small version number differences.

A.1.3 Manual Review and Missing Benchmarks

Despite sophisticated automated matching, many models require manual verification due to genuine ambiguity or high stakes for analysis accuracy. We implement a systematic manual review process with multiple checkpoints. After running the five-stage algorithmic matching, we manually review all matches generated by stages 4 and 5 (substring containment and Levenshtein distance) due to their higher error rates. We also manually verify matches for all models in the top quartile of OpenRouter usage, as errors in matching high-volume models would disproportionately impact our analyses. For models with multiple potential matches (such as when a single OpenRouter model could map to several Hugging Face variants), we review documentation, release dates, and benchmark scores to identify the most appropriate match. This manual review process involves cross-referencing model cards on Hugging Face, technical documentation from model providers, and release announcements to verify model identity.

Not all models on each platform successfully match to models on other platforms. OpenRouter lists many specialized or niche models that have never been formally benchmarked on Artificial Analysis or LM Arena. Conversely, some models evaluated by benchmark platforms are no longer available on OpenRouter or were never served there. Some models appear on Hugging Face primarily as base models for fine-tuning rather than for inference, and thus do not appear in our usage data. For our main analyses, we focus on models on OpenRouter that successfully match across at least one other data sources (OpenRouter for usage plus at least one benchmark source for capabilities). This restriction ensures we can analyze the relationship between model usage and measured capabilities, which is central to our research questions. We document the share of total usage covered by matched versus unmatched models in our results section to assess whether our

matched sample represents the bulk of market activity. We discuss the effect of missing benchmarks in A.2.

A.2 Missing Benchmark Data and Selection into the Regression Sample

Our usage regressions in Section 3.2.1 relate log July 2025 token usage to log price, an open-source indicator, and benchmark performance measures. These benchmarks are only available for a subset of models, raising the concern that restricting attention to models with benchmark data may bias the estimated relationship between openness and usage. Figure A1 plots all 260 models with non-missing price and July usage, highlighting which ones have scores on each benchmark. Two facts stand out. First, benchmark coverage is highly correlated with usage: virtually all high-usage models have at least one score on GPQA, MMLU-Pro, or LM Arena, whereas missing benchmark data is concentrated among models with very low token volumes. Second, coverage is similar for open and closed models within the economically important range of token usage and prices; unbenchmarked models are predominantly small or niche models that account for a negligible fraction of total tokens.

Conceptually, selection into the "benchmarked" sample is far from random: models are more likely to be evaluated if they are widely used or otherwise salient. Under the plausible assumption that unbenchmarked models would tend to score worse than benchmarked models on these generic tests, and that they also have substantially lower usage (as Figure A1 suggests), omitting them has two main consequences for our regressions. First, dropping low-usage, low-performance models is equivalent to estimating the usage–performance relationship on a truncated upper tail of the token distribution. In simple linear models, such truncation tends to flatten the estimated slope: we understate how strongly usage responds to benchmark performance. Second, to the extent that low-usage models are disproportionately open source, the benchmarked subsample over-represents relatively strong open models. As a result, the negative "open-source penalty" we estimate—coefficients ranging from approximately -1.0 to -2.1 (equivalent to roughly 100–210 log points) representing a 63% to 88% reduction in usage for open models relative to closed models, conditional on price and benchmarks—should be interpreted as a lower bound in magnitude for the universe of all models. Including the many small, low-usage open models that never get benchmarked would, if anything, make the estimated underutilization of open models larger.

These interpretations are supported by two pieces of evidence. First, the open-source coefficient remains large and negative in specifications that use the full sample of models and control only for price (specification (2) in Table A3), i.e., in a regression that is not affected by benchmark missingness. Second, coefficients on openness are stable in sign and similar in magnitude across specifications that use different benchmark subsets (specifications (3)–(8) in Table A3 and Table A4), despite substantial changes in sample size. Together, these patterns suggest that sample selection due to missing benchmarks does not drive our main conclusion that open models are systematically underutilized relative to closed models.

Finally, our observable domination and consumer savings calculations in Section 3.2.3 are, by construction, computed only over models with valid benchmark scores. Given that unbenchmarked models are overwhelmingly low-usage and cannot serve as dominating alternatives in price–performance space, this restriction makes our estimates conservative. Any additional mis-

allocation involving unbenchmarked models—e.g., closed models that are dominated by unbenchmarked open models—would increase the implied scope for efficiency gains beyond the magnitudes we report.

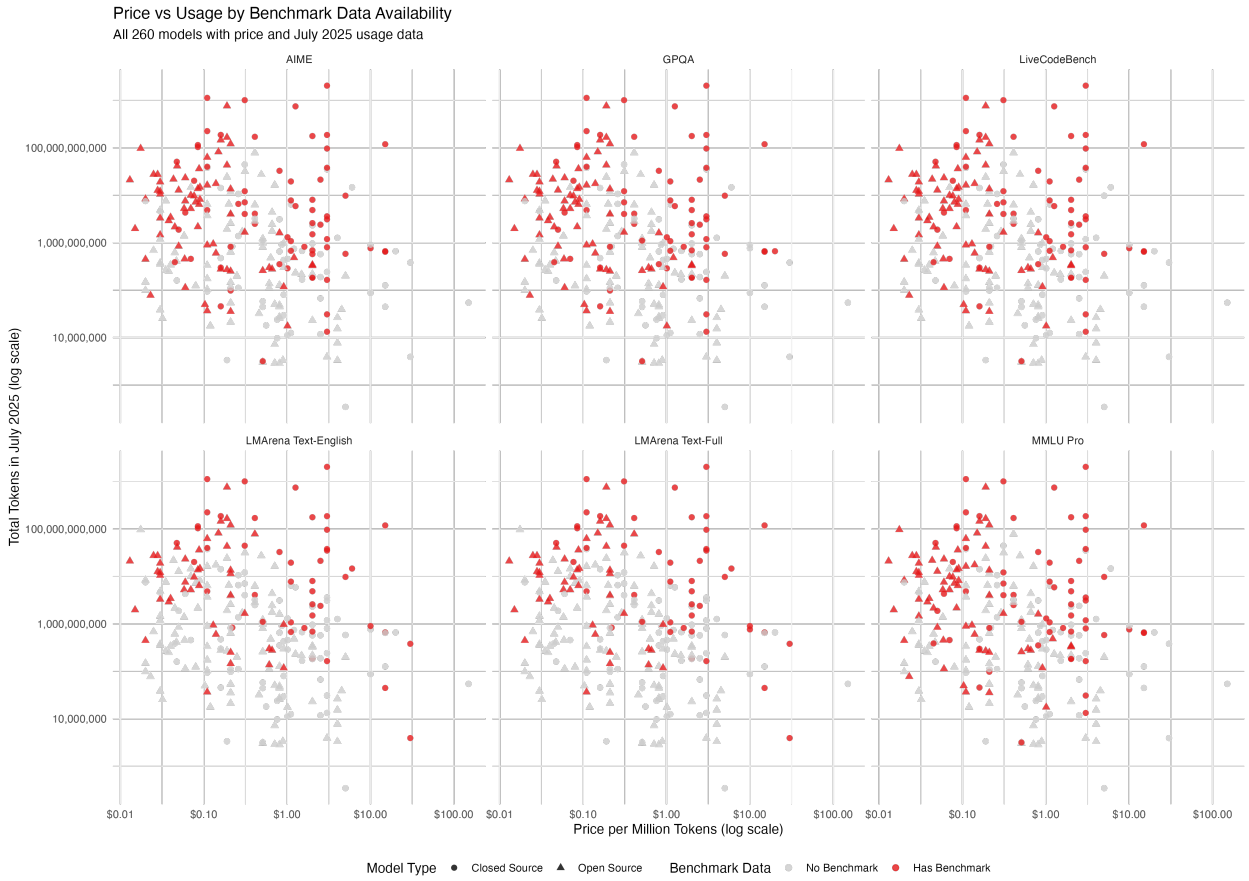

Figure A1: Benchmark Coverage by Model Usage and Price (July 2025). This figure plots all 260 models with non-missing price and July 2025 usage data, showing which models have benchmark scores available for GPQA, MMLU-Pro, and LM Arena. Panels show coverage for each benchmark separately. Benchmark coverage is highly correlated with usage: virtually all high-usage models (top of vertical axis) have benchmark scores, while missing data is concentrated among low-usage models. Coverage is similar for open and closed models within the economically important range of usage and prices.

A.3 Modeling the Growth of API Inference Expenditure (2024–2025)

To approximate the growth of spending in the AI LLM API market from 2024 to 2025, we model the spending rate as a continuous function $f(t)$, where t is measured in years since the start of 2024 using the observed variables from the Menlo Ventures report (Tully et al., 2025) — an estimated \$3.5B market size in 2024, and \$8.4B in the first half of 2025. The total expenditure over a time interval is given by the integral of $f(t)$.

A.3.1 Exponential Baseline Fit

We begin with an exponential model

$$f(t) = Ae^{kt},$$

and use the known integrals from the Menlo Venture report

$$\int_0^1 f(t)\,dt = \$3.5 \text{ B}, \qquad \int_1^{1.5} f(t)\,dt = \$8.4 \text{ B}.$$

Solving these equations gives

$$A \approx 0.896, \qquad k \approx 2.304.$$

The model implies a sharp acceleration, with the integral over the second half of 2025

$$\int_{1.5}^2 f(t)\,dt \approx \$26.7 \text{ B}.$$

A.3.2 Alternative Functional Forms

For comparison, two other models were fit to satisfy the same two integral constraints:

- **Linear:** $f(t) = \alpha + \beta t$, giving $\alpha \approx -5.37$, $\beta \approx 17.73$.
- **Power-law:** $f(t) = at^p$, giving $a \approx 10.56$, $p \approx 2.02$.

The resulting 2025 expenditure projections (in billions) are summarized below:

Model	H1 2025	H2 2025	Total 2025
Exponential	\$8.4	\$26.7	\$35.1
Linear	\$8.4	\$12.8	\$21.2
Power-law	\$8.4	\$16.5	\$24.9

A.4 Supplementary Figures

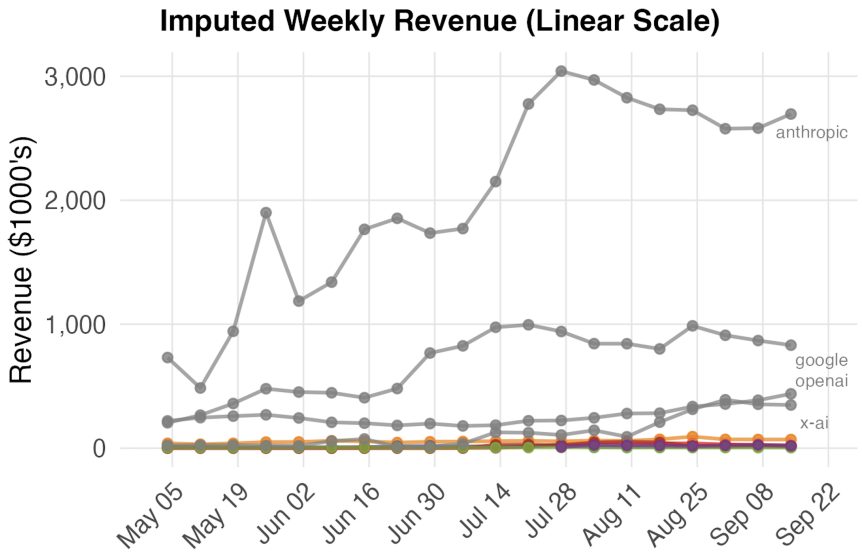

Figure A2: Weekly Revenue by Model Provider (Linear Scale). The same data as Figure 2 shown on a linear scale, more clearly illustrating the magnitude of revenue differences between closed and open model providers.

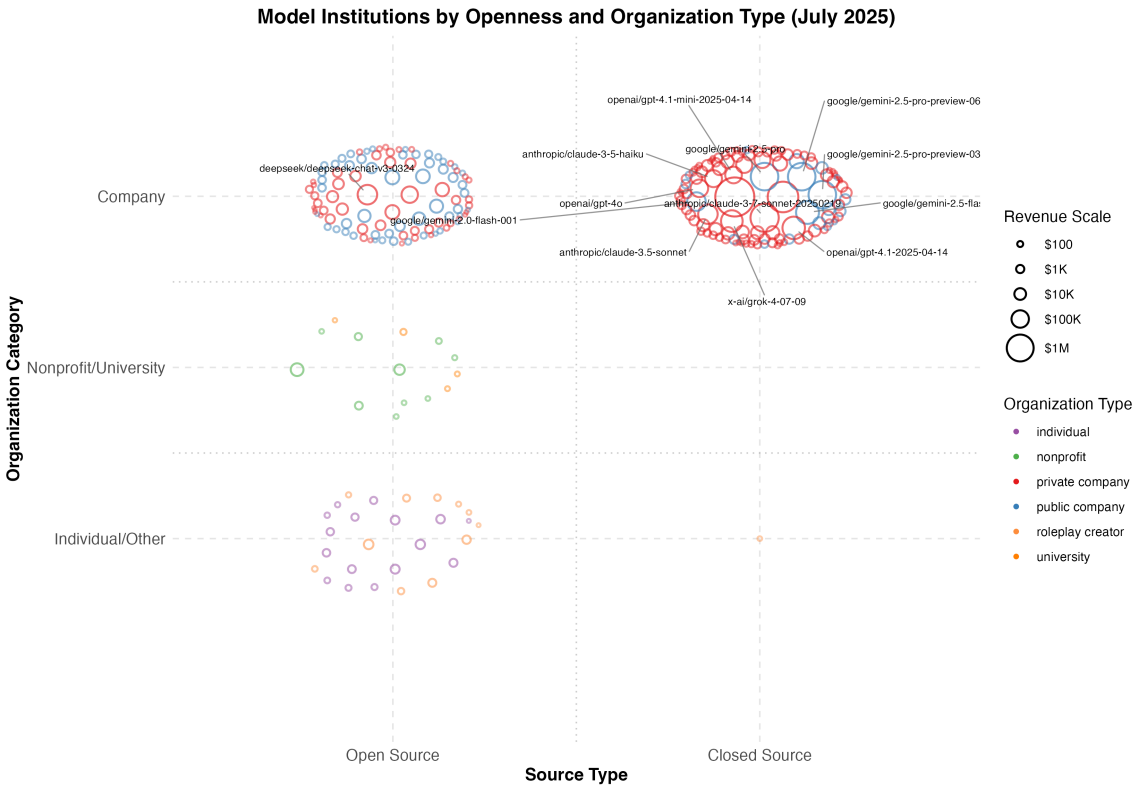

Figure A3: Model Institutions by Openness and Organization Type (July 2025). This figure displays the organizational origins of AI models, categorized by whether they are open source or closed source and by organization type (company, nonprofit/university, or individual/other). Bubble size represents revenue scale. The plot demonstrates that all high-usage models with significant revenue are created by companies, with both open and closed models predominantly originating from corporate entities rather than academic institutions or individual developers.

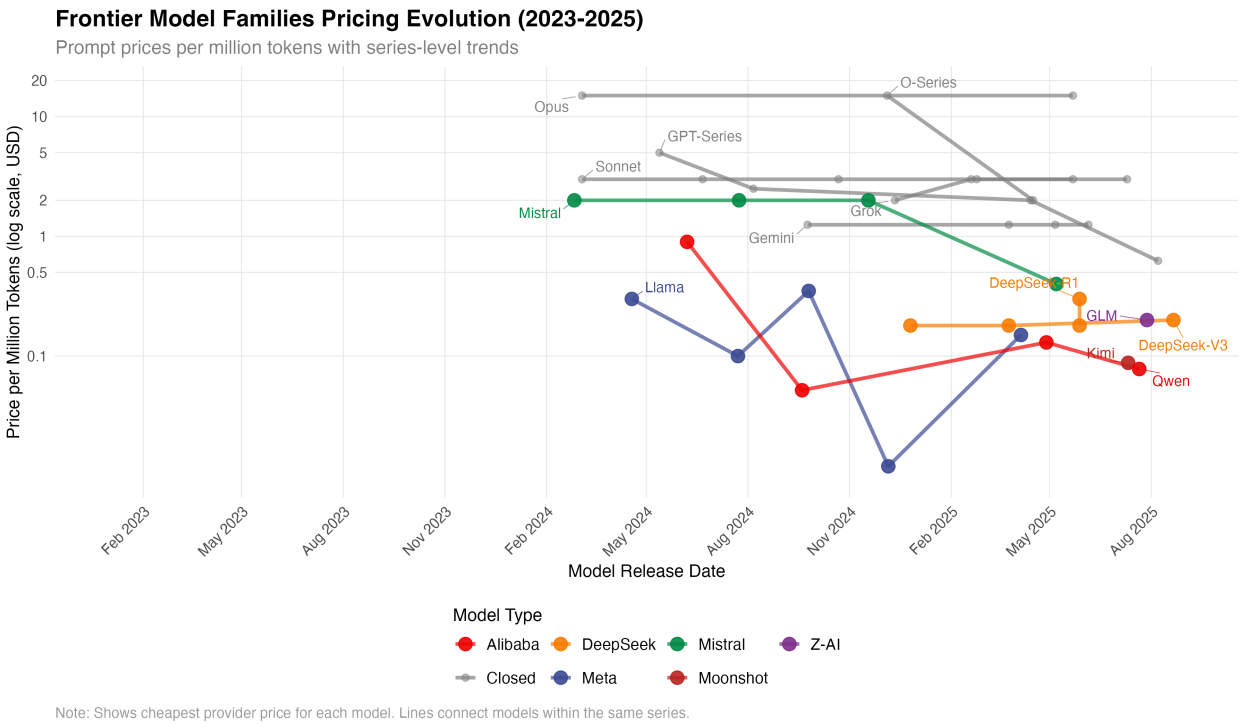

Figure A4: Frontier Model Families Pricing Evolution (2023-2025). Prompt prices per million tokens for leading model families over time, shown on a log scale. This figure displays pricing trends for the most prominent model families only, but is indicative of the relative stability of prices within model families over the analysis period. Lines connect models within the same series, with each point representing a model release. The cheapest provider price is shown for each model.

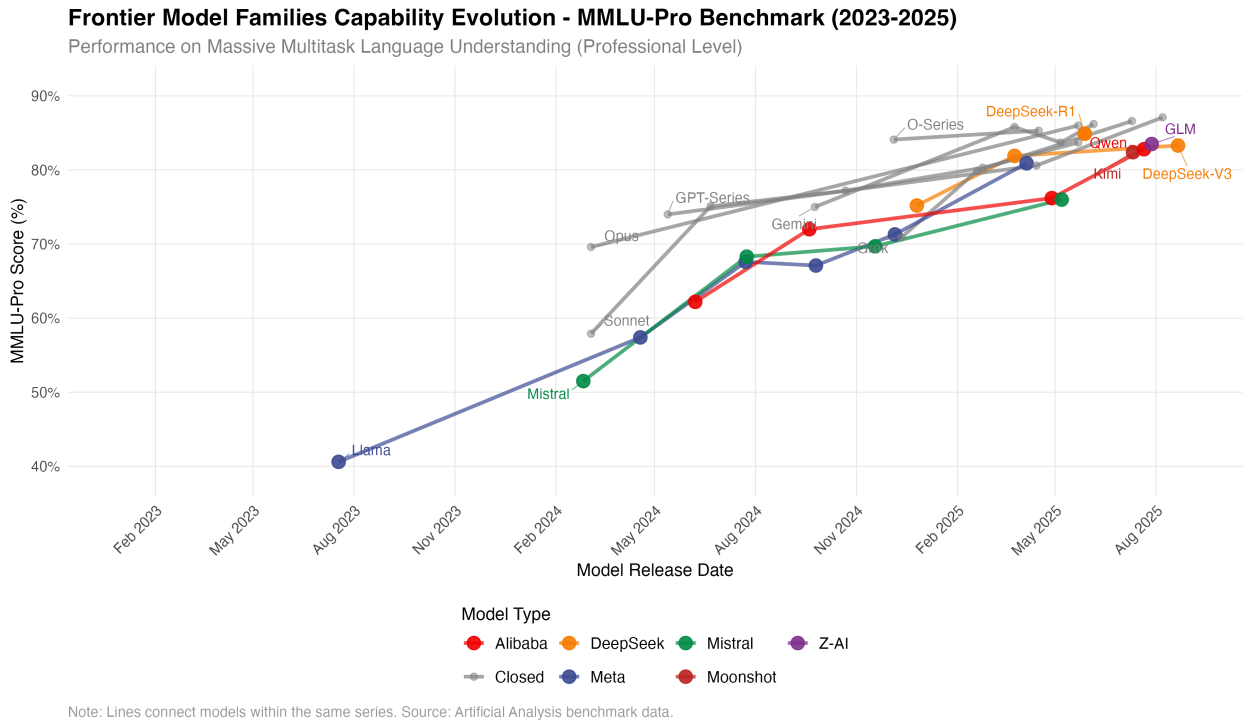

Figure A5: Frontier Model Families MMLU Pro Performance Evolution (2023-2025). MMLU Pro scores for leading model families over time, showing the evolution of both closed (red) and open (blue/orange) models. Lines connect models within the same family, with each point representing a model release. The figure demonstrates that the pattern of open models rapidly catching up to closed models observed for GPQA (see Figure 4) is robust to the choice of benchmark, with similar dynamics visible for MMLU Pro performance.

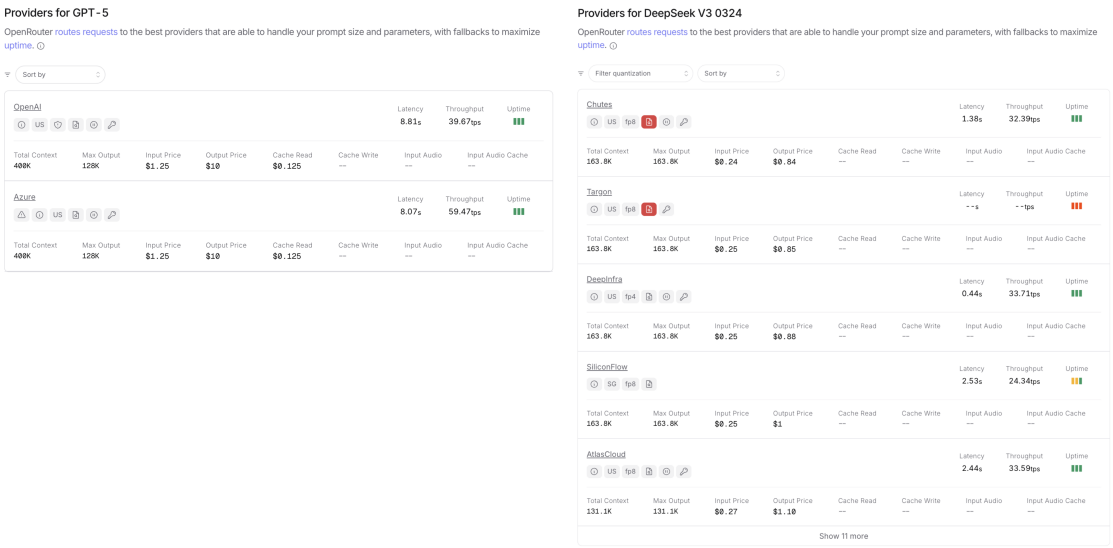

Figure A6: Example Provider Listings on OpenRouter: Closed vs Open Models. This figure shows OpenRouter's provider listings for two models as of September 2025. The left panel shows GPT-4, a closed model available exclusively through OpenAI. The right panel shows DeepSeek-V3, an open model available through six different inference providers (Chutes, Targon, DeepInfra, SiliconFlow, AtlasCloud, and Novita). Multiple providers compete on dimensions including price, latency, throughput, and service quality, creating competitive pressure that drives down open model pricing.

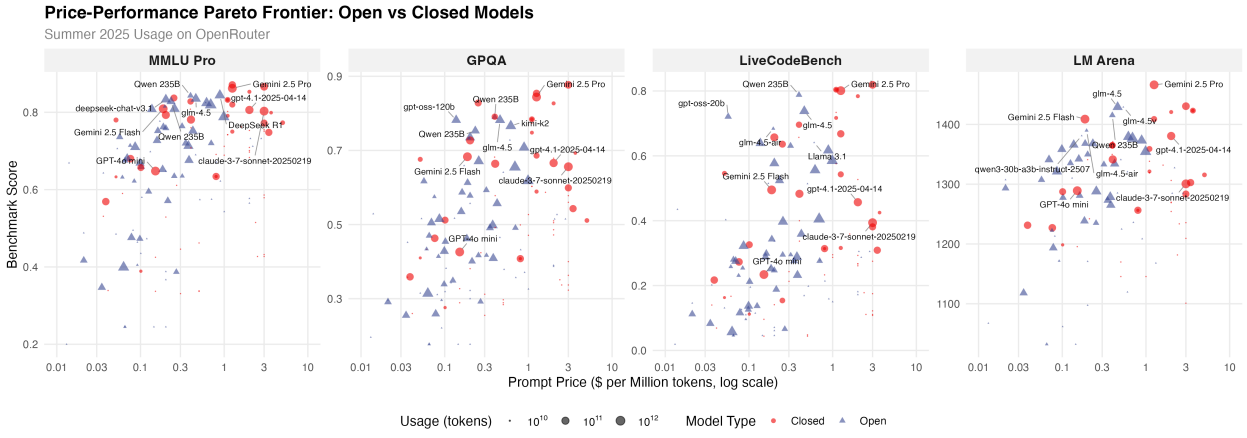

Figure A7: Price-Performance Frontier: Open vs Closed Models (Full Benchmarks). This figure extends Figure 8 by showing the price-performance relationship across all four key benchmarks: MMLU Pro, GPQA, LiveCodeBench, and LM Arena text scores. Models are plotted with bubble sizes proportional to usage (total tokens processed in summer 2025). The pattern of observable domination is consistent across benchmarks: several high-performing, low-cost open source models (predominantly Chinese models such as Qwen 235B, DeepSeek-V3, and GLM-4.5) dominate popular closed models from Western providers (such as GPT-4.1, Claude Sonnet variants, and Gemini Flash variants) on both price and performance dimensions, yet receive substantially lower usage.

A.5 Supplementary Tables

Table A1: Unrealized Value of Open Models: Simulation Results (Dominated Improvements)

Method	Benchmark	% Dominated Closed Models	% Dominated Closed Tokens	Average Price	Average Imputed Price	Effective Price Change (%)	Average Benchmark	Average Imputed Benchmark	Effective Benchmark Change (%)
Best Performance	GPQA	90.4%	71.1%	$1.69	$0.65	-61.9%	0.68	0.80	16.9%
Best Performance	LM Arena	93.9%	76.5%	$1.79	$0.56	-68.9%	1371.59	1425.56	3.9%
Best Performance	LiveCodeBench	90.0%	86.7%	$1.69	$0.40	-76.1%	0.51	0.76	51.0%
Best Performance	MMLU Pro	92.7%	83.5%	$1.69	$0.72	-57.4%	0.80	0.84	5.2%
Best Performance	**Method Average**	91.7%	79.4%	$1.72	$0.58	-66.1%			19.3%
Best Value	GPQA	90.4%	71.1%	$1.69	$0.53	-68.9%	0.68	0.78	14.8%
Best Value	LM Arena	93.9%	76.5%	$1.79	$0.53	-70.6%	1371.59	1417.75	3.4%
Best Value	LiveCodeBench	90.0%	86.7%	$1.69	$0.34	-80.0%	0.51	0.73	43.6%
Best Value	MMLU Pro	92.7%	83.5%	$1.69	$0.58	-65.5%	0.80	0.84	4.3%
Best Value	**Method Average**	91.7%	79.4%	$1.72	$0.49	-71.3%			16.5%
Lowest Price	GPQA	90.4%	71.1%	$1.69	$0.48	-71.7%	0.68	0.76	10.9%
Lowest Price	LM Arena	93.9%	76.5%	$1.79	$0.45	-74.7%	1371.59	1382.48	0.8%
Lowest Price	LiveCodeBench	90.0%	86.7%	$1.69	$0.26	-84.6%	0.51	0.58	14.6%
Lowest Price	MMLU Pro	92.7%	83.5%	$1.69	$0.54	-68.3%	0.80	0.82	1.8%
Lowest Price	**Method Average**	91.7%	79.4%	$1.72	$0.43	-74.8%			7.0%
Overall Average	Overall Average	91.7%	79.4%	$1.72	$0.50	-70.7%			14.3%

Note: This table presents simulation results of the unrealized value analysis for closed source models with dominated improvements from open alternatives. Data covers June-August 2025 usage on OpenRouter (14 weeks). % Dominated Models: percentage of unique closed models with a superior open alternative. % Dominated Tokens: percentage of tokens from closed models with superior alternatives. Prices are per million tokens. All averages are token-weighted. Benchmark differences show open minus closed scores; positive values indicate open models outperform closed models.

Table A2: Estimating the Size of the Total AI Inference Market

Time Period	Reference Method	Reference Value	OR Computed Value	OR Fraction	OR Annual Revenue	Estimated Annual Market Size
May-June 2025	Google Tokens (Demis Tweet)	1,460T	6.2T	0.43%	$182M	$42.7B
May-June 2025	OpenAI API Revenue (Epoch AI Estimates)	$ 0.3B	$ 1M	0.30%	$182M	$60.4B
Jan-June 2025	Menlo Ventures H1 Estimate	$ 8.6B	$ 91M	1.06%	$182M	$35.1B

Note: This table presents aggregated time period extrapolations. Google Tokens reference is from Demis Hassabis tweet. OpenAI API Revenue estimates use Epoch AI total revenue data ($10B ARR in May, $12B ARR in July) multiplied by an estimated 20% API fraction based on industry reports. Menlo Ventures H1 estimate ($8.6B) represents their H1 2025 generative AI market size projection. OR Fraction shows what percentage of the reference value is observed on OpenRouter. OR Annual Revenue shows OpenRouter revenue scaled to a full year ($182M). Estimated Annual Market Size shows the extrapolated annual market size; for Menlo, the annual estimate is $35.1B (see A.3).

Table A3: Model Usage and Price Regressions: Main Effects. This table reports regression results for log token usage and log price per million tokens on model openness and standardized benchmark scores for July 2025 data. The open source variable is centered (-0.5 for closed, 0.5 for open). Models (1)-(2) include only openness and price; models (3)-(4) add Artificial Analysis benchmarks (AIME, GPQA, LiveCodeBench, MMLU Pro); models (5)-(6) add LMArena scores; models (7)-(8) include all available benchmarks. The consistently negative coefficient on open source in usage regressions indicates that open models receive lower usage than closed models even after controlling for price and performance. Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Dependent Variables:	Log(Price)	Log(Tokens)	Log(Price)	Log(Tokens)	Log(Price)	Log(Tokens)	Log(Price)	Log(Tokens)
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Open Source	-1.467***	-1.267***	-1.361***	-1.135**	-2.120***	-1.358**	-1.653***	-2.099***
	(0.1877)	(0.3849)	(0.2320)	(0.4867)	(0.3227)	(0.5931)	(0.2896)	(0.6391)
Log(Price)		-0.5773***		-0.9005***		-0.8141***		-1.090***
		(0.1141)		(0.1683)		(0.1569)		(0.2138)
AIME (std)			-0.8620***	-1.253**			-0.9091**	-1.810**
			(0.2800)	(0.5382)			(0.4225)	(0.8010)
GPQA (std)			0.3800	1.054			0.2886	0.6742
			(0.4284)	(0.7955)			(0.5411)	(0.9971)
LiveCodeBench (std)			0.6738*	0.6360			0.8551*	0.8771
			(0.3850)	(0.7217)			(0.4735)	(0.8898)
MMLU Pro (std)			0.2648	1.140**			0.7279*	0.9953
			(0.2775)	(0.5156)			(0.4044)	(0.7599)
LMArena Text-Full (std)					-0.7486	1.460	-2.104**	-0.8895
					(1.128)	(1.720)	(1.000)	(1.894)
LMArena Text-English (std)					0.9054	0.1093	1.658*	1.729
					(1.123)	(1.714)	(0.9743)	(1.827)
Constant	-0.8078***	20.38***	-1.183***	21.20***	-0.8780***	21.93***	-1.336***	21.58***
	(0.0938)	(0.1963)	(0.1084)	(0.2826)	(0.1443)	(0.2592)	(0.1296)	(0.3720)
Fit statistics								
Observations	264	264	127	127	98	98	82	82
R^2	0.18916	0.09394	0.45171	0.36181	0.35938	0.42730	0.56336	0.42562
Adjusted R^2	0.18606	0.08699	0.42906	0.32990	0.33894	0.40267	0.52206	0.36268

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A4: Model Usage and Price Regressions: Interaction Effects. This table extends Table A3 by including interaction terms between the open source indicator and benchmark scores, as well as between open source and log price. These interactions test whether the relationship between capabilities and usage differs systematically for open versus closed models. The positive interactions between open source and AIME scores suggest that high-performing open models see relatively higher usage, though the negative main effect of openness dominates. Models (1)-(2) include Artificial Analysis benchmarks; models (3)-(4) include LMArena scores; models (5)-(6) include all available benchmarks. Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Dependent Variables:	Log(Price)	Log(Tokens)	Log(Price)	Log(Tokens)	Log(Price)	Log(Tokens)
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Open Source	-1.291***	-0.9280	-2.160***	-1.027	-1.702***	-1.303
	(0.2305)	(0.6550)	(0.3221)	(0.6930)	(0.2986)	(0.9325)
AIME (std)	-1.079***	-1.388**			-0.8799**	-1.882**
	(0.2918)	(0.5623)			(0.4095)	(0.7917)
GPQA (std)	0.7440	1.135			0.6108	0.5704
	(0.4580)	(0.8398)			(0.5407)	(1.044)
LiveCodeBench (std)	0.6495*	0.5151			0.6212	0.8125
	(0.3877)	(0.7088)			(0.4651)	(0.8670)
MMLU Pro (std)	0.1126	1.397***			0.6328	1.303*
	(0.2910)	(0.5274)			(0.3971)	(0.7712)
Open Source × AIME	1.275**	1.418			1.546*	3.268**
	(0.5835)	(1.125)			(0.8189)	(1.583)
Open Source × GPQA	-1.301	0.9793			-2.719**	-0.8421
	(0.9161)	(1.680)			(1.081)	(2.088)
Open Source × LiveCodeBench	-0.7799	-2.148			-0.3114	-2.187
	(0.7755)	(1.418)			(0.9301)	(1.734)
Open Source × MMLU Pro	0.6341	-1.718			1.586**	-0.5510
	(0.5819)	(1.055)			(0.7943)	(1.542)
Log(Price)		-0.9732***		-0.7348***		-1.093***
		(0.1675)		(0.1657)		(0.2260)
Log(Price) × Open Source		0.1897		0.0388		0.2335
		(0.3350)		(0.3314)		(0.4520)
LMArena Text-Full (std)			-0.9418	2.383	-2.747***	-0.2418
			(1.157)	(1.762)	(1.013)	(1.980)
LMArena Text-English (std)			1.027	-0.7071	2.161**	1.259
			(1.152)	(1.752)	(0.9745)	(1.875)
Open Source × LMArena Text-Full			3.962*	-5.811	3.223	-2.623
			(2.314)	(3.525)	(2.026)	(3.961)
Open Source × LMArena Text-English			-3.668	4.750	-2.969	1.705
			(2.304)	(3.503)	(1.949)	(3.750)
Constant	-1.301***	20.99***	-0.7417***	21.79***	-1.306***	21.31***
	(0.1153)	(0.3275)	(0.1611)	(0.3465)	(0.1493)	(0.4663)
Fit statistics						
Observations	127	127	98	98	82	82
R^2	0.48466	0.43703	0.38343	0.47083	0.62707	0.51683
Adjusted R^2	0.44502	0.38318	0.34992	0.42967	0.55577	0.40702

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*