The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs *

Mert Demirer† Andrey Fradkin‡ Nadav Tadelis§ Sida Peng¶

December 12, 2025

Abstract

We document six facts about the structure and dynamics of the LLM market using API usage data from OpenRouter and Microsoft Azure. First, we show rapid growth in the number of models, creators, and inference providers, driven by open-source entrants. Second, we show price declines and persistent price heterogeneity across and within intelligence tiers, with open-source models being 90% cheaper than comparable closed-source models of the same intelligence. Third, we document market dynamism, with frequent turnover among leading models and creators. Fourth, we present evidence of horizontal and vertical differentiation, with no single model dominating across use cases, and demand for intelligence varying widely across applications. Fifth, we estimate preliminary short-run price elasticities just above one, suggesting limited scope for Jevons-Paradox effects. Finally, we show that although the share of firms that use multiple models increased over time, most firms concentrate their use on a single model, consistent with experimentation rather than persistent reliance on multiple models.

*We thank Alessandro Bonatti, Seth Benzell, James Brand, Erik Brynjolfsson, Tom Cunningham, Dean Eckles, Garrett Johnson, Daniel Rock, and seminar participants at the Stanford Digital Economy Lab, MIT Sloan, and MIT FutureTech. We thank Marilyn Pareboom for exceptional research assistance. The views of this paper do not necessarily reflect the views of Microsoft or Amazon.

†MIT Sloan, mdemirer@mit.edu

‡Boston University & Amazon Inc., fradkin@bu.edu

§Microsoft, nadavtadelis@microsoft.com

¶Microsoft, sidpeng@microsoft.com

1 Introduction

Artificial intelligence is a general-purpose technology that holds the promise of increasing productivity, enabling new products, and transforming the economy. Given the importance of this technology, it is critical for businesses, policymakers, and researchers to measure how AI is spreading throughout the economy, as well as the market structure of AI. In this paper, we document new facts about the supply, demand, and pricing of AI models by using data from two of the largest marketplaces for AI APIs, OpenRouter and Microsoft Azure. Of particular interest, we provide initial estimates of demand for tokens, showing that short-run elasticities in the API market are unlikely to justify the Jevons Paradox, in which prices falling cause total quantity demanded to increase.

We first document the rapid growth of LLMs, their creators, and companies providing LLM inference for open-source models, such as DeepInfra, Fireworks, and Groq. The number of distinct models available has grown from just over 253 to over 651 between January 2025 and December 2025. Simultaneously, the number of creators of models has almost doubled from 43 at the beginning of the year to 85 by the beginning of December. Most of these entrants specialize in open-source models, whereas the number of closed-source model creators remains relatively stable. The number of inference providers, primarily serving open-source models, has grown even more rapidly, increasing from 27 in early 2025 to 90 by late 2025. This has led to increasing competition in the provision of open-source models, with some models being served by more than 20 different providers. Concurrent with these trends is the growth of OpenRouter itself, which has served over 100 trillion tokens this year.

Second, we document trends in LLM pricing by showing both a significant decline over time and substantial heterogeneity across and within intelligence tiers. Models that were state-of-the-art in 2023 have experienced a price decline of approximately 1000 times, with similarly pronounced deflationary trends at other intelligence levels. The average price paid per token has remained relatively constant, consistent with demand for superior intelligence. At the same time, there is substantial variation in the prices of models that, according to benchmarks, are similar in intelligence. Of particular note is that open-source models are approximately 90% cheaper than closed-source models, conditional on the same level of intelligence. Nonetheless, the share of tokens consumed from open-source models remains consistently below 30%, suggesting meaningful differentiation between open and closed-source models not captured by intelligence measures.

Third, we document substantial market dynamism, with frequent fluctuations in market

[0]Throughout the paper, we use the terms AI and LLMs interchangeably, recognizing that the architecture of the models may change over time.

shares across both models and creators. At the model level, the leading model typically holds the top position for only a few months before being displaced, and the top 10 models today accounted for just 20% of market share four months ago and did not even exist ten months ago. Perhaps more importantly, we observe similar churn at the *creator* level: over the course of the year, models from Anthropic, Google, and xAI each held the top market share on OpenRouter at different points. Open-source models gained market share through August 2025, but their overall market share has declined since then.

Fourth, we find indirect evidence of both horizontal and vertical differentiation between models. No single model dominates across use cases (i.e., programming and marketing), and models with comparable intelligence levels often serve as the top choice for different applications. Moreover, demand for intelligence varies substantially across use cases: categories such as programming rely on models close to the frontier, whereas role-play and translation are dominated by models with considerably lower intelligence. As a result, overall demand is not concentrated on frontier models on both OpenRouter and Azure: tokens even at the 90th percentile of the intelligence distribution remain well below the highest available intelligence levels. This pattern has two implications. First, there is meaningful scope for competition not only among the most capable models but also among mid-tier and lower-intelligence models. Second, for many use cases, the incremental gains in intelligence offered by frontier models do not appear to justify their higher prices.

As supporting evidence for the horizontal and vertical differentiation, we also examine how demand responds to new model entries on OpenRouter and Azure and find that distinct model entries yield divergent substitution patterns. Anthropic models, particularly in the Sonnet family, primarily steal demand from previous Sonnet models but do not substantially affect demand for other models on OpenRouter. This is consistent with horizontal differentiation between Anthropic models and those of other providers, and with vertical differentiation within Anthropic models. In contrast, models such as Google's Gemini Flash 2.0 and xAI's Grok Code Fast 1 are successful but do not cause immediate substitution from other models.

Fifth, we examine the determinants of token demand. We first provide a simple framework for thinking about token demand, distinguishing between the short run and the long run and between models and within models. We then estimate regressions of tokens on prices, performance metrics such as latency, and other model characteristics. The primary identification challenge in these regressions is price endogeneity, which we cannot fully eliminate. Our most credible specification uses day-to-day variation in prices across providers within the model. In our preferred estimates, we find price elasticities just above one for model–provider combinations, suggesting that short-run Jevons' Paradox is unlikely to operate at the market level.

Lastly, we examine the extent to which firms use multiple models simultaneously (multihoming). We find that, in any given month, more than 50% of firms rely on a single model, though the share using multiple models has grown steadily over time—from about 25% to 50% between mid-2023 and mid-2025. However, when we focus on the intensive margin among multi-homers, we find that most firms allocate more than 90% of their total usage to a single model. This pattern suggests that, for most firms, multi-homing reflects experimentation rather than persistent, task-specific reliance on multiple models.

Our analysis is primarily based on a scraped dataset from OpenRouter. We also confirm many of the findings using aggregated firm-level API usage data from Microsoft Azure. OpenRouter provides an API that enables app developers and other users to manage interactions with a variety of LLMs, including tools for routing API calls across models and providers based on price and latency. As a marketplace, OpenRouter provides an excellent setting for studying changes in demand for LLMs by app developers and their end users. Our OpenRouter data allows us to observe the tokens called by the model, date, and, in some cases, the application and category of use. In addition, we observe data on model providers. We augment the above data with data on model benchmarks from Artificial Analysis.

1.1 Related Literature

The academic literature on AI has primarily focused on three aspects of LLMs: (i) analyses of occupational exposure to AI (Brynjolfsson et al., 2018; Felten et al., 2018, 2021, 2023; Eloundou et al., 2024; Handa et al., 2025; Shao et al., 2025; Hampole et al., 2025; Demirer et al., 2025); (ii) macroeconomic modeling of AI (Acemoglu and Restrepo, 2018, 2019; Acemoglu, 2024; Autor and Thompson, 2025); and (iii) empirical evaluations of AI's productivity effects (Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Brynjolfsson et al., 2025; Cui et al., 2025). The exposure literature examines actual LLM prompts or task characteristics to quantify the extent to which occupations are affected by AI. The macroeconomic literature incorporates AI into economic models to assess its aggregate impacts. Finally, the productivity literature relies on experiments or observational data to measure how AI adoption influences productivity within specific occupations or tasks.

Our paper differs from the existing literature by focusing directly on the LLM market itself and the enterprise use of LLMs. It supercedes a preliminary analysis in Fradkin (2025). To the best of our knowledge, there is no comprehensive academic work on the structure and dynamics of the LLM market. Nagle and Yue (2025), in simultaneous work, investigates the role of open-source models, and Aubakirova and Midha (2025) provides an industry-oriented report using OpenRouter data. Additional existing evidence has largely come from industry trackers such as Artificial Analysis and EpochAI, which provide pricing and performance

comparisons across models. On the enterprise side, prior research has relied almost exclusively on survey evidence, both in the academic literature on AI adoption (Bick et al., 2024; Humlum and Vestergaard, 2025) and in industry reports (Deloitte, 2024; McKinsey, 2024; Eurostat, 2025; Stanford Institute for Human-Centered Artificial Intelligence, 2025). A smaller set of sources further analyzes demand for LLM models, again based on surveys (Kong Inc., 2024; Andreessen Horowitz, 2025; Menlo Ventures, 2025). We contribute to this literature by providing the first comprehensive, non-survey-based dataset on LLM demand.

One advantage of our analysis is that we examine a broad set of models sourced from multiple providers. Most AI labs produce research based solely on their own proprietary datasets (e.g., OpenAI and Anthropic; see Chatterji et al. 2025, OpenAI 2025, and Handa et al. 2025), which makes it difficult to systematically compare models across providers. We circumvent this limitation by drawing on two distinct sources—OpenRouter and Microsoft Azure. Moreover, leveraging a large aggregator such as OpenRouter allows us to compare and evaluate models from numerous providers within a unified empirical framework.

2 The Market for LLMs

We consider the market for LLM inference sold via APIs, measured in metered token usage. This market is structured in a vertical supply chain, with some players occupying multiple positions. First, there are LLM creators such as Anthropic, OpenAI, and Google, who train models, or in some cases modify open-source models. Second, there are inference providers, such as Azure, Cerebras, Google Cloud, and Together AI. These companies operate compute clusters that serve a subset of models to their users. Importantly, the same LLM may be served by multiple inference providers. Lastly, there are aggregators such as OpenRouter that operate a marketplace for tokens across providers and models.[1]

Our focus is on LLM usage for business purposes rather than direct consumer demand via messaging interfaces such as ChatGPT. This is a substantial subset of the broader AI market. Enterprises can use AI services in a variety of ways, including training or fine-tuning models, purchasing software on a seat-based license, procuring consulting services, or making API calls to an existing model, either directly or through an Integrated Development Environment such as Cline or Cursor. We focus only on API access, since it is by far the most accessible and the most common mode of adoption for enterprises.[2]

API access enables a firm's internal systems to send text or multi-modal requests to a model hosted by a provider. The request, often called a "prompt," is transmitted over the

[1] As far as we know, OpenRouter is the only aggregator with substantial enterprise traction. In the consumer market, apps such as LMSys, Poe, and Yupp allow users to try a variety of LLMs.

[2] Menlo Ventures estimates that total enterprise spending on API access for Generative AI increased from $3.5 billion at the start of 2025 to $8.4 billion by mid-year.

internet through a standardized interface, and the model returns a generated response that the firm can directly integrate into its own applications.[3] This design treats the model as infrastructure rather than a standalone product: firms do not need to manage training or deployment themselves, but instead rent intelligence on demand, scaling their usage up or down depending on their needs.

We now describe each type of actor in this market in greater detail.

2.1 LLM Creators

Creators are the entities that develop and train LLMs. Their main contribution is the design of the model architecture, the assembly of large-scale training datasets, and the execution of computationally intensive training runs. Many creators keep their models closed-source, meaning that access is only possible through commercial arrangements with providers. In some cases, creators also act as their own providers—for example, Google and Microsoft develop models internally and offer them directly only through their cloud platform.

At other times, creators distribute their models through multiple providers, making them accessible across different cloud ecosystems. For instance, Anthropic develops the Claude family of models and makes them available through Microsoft Azure, Google Cloud, Amazon Web Services, and on its own platform. This multi-provider strategy expands reach and adoption while still allowing the creator to control access terms and pricing.

For closed-source models, competition among creators is limited by the data, computational resources, and expertise required to train frontier-scale LLMs. This concentration gives creators significant influence over the market's trajectory and shapes the bargaining relationships among creators, providers, and downstream enterprises.

Some firms train models but release them under an open-source license. This has, to date, been the strategy of Meta with its Llama model family, as well as DeepSeek, Moonshot AI, and others. In the summer of 2025, OpenAI released its first open-source LLM. An interesting phenomenon regarding open-source models is that they can be modified by others. For example, models can be fine-tuned to perform better at certain tasks. They can also be 'distilled', which means that a smaller model can be made to emulate the outputs of a larger model. We observe both types of models being used in the market.

2.2 LLM Inference Providers

Inference occurs when models return completion tokens in response to a prompt. Given the large size of these models, specialized large-scale compute systems are needed, and inference providers design and offer these systems. To serve the inference market, companies require

[3]LLMs differ in their context window, which is the size of the prompt that they are able to ingest.

computing resources, model weights, and orchestration software, often developed in-house. The available compute is determined by the processors an inference firm has and the energy required to run them. There is ongoing innovation in optimizing compute systems for inference.

Inference services are provided by a variety of firms. The three major cloud platforms—Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)—offer inference via API. These companies serve a variety of models, and also have special relationships with certain model creators, which allow them to serve closed-source models. Specifically, AWS and GCP serve Anthropic's models, while Azure serves OpenAI's models.[4]

Model creators such as Anthropic and OpenAI also operate their own API endpoints. These endpoints likely use compute from the major cloud platforms but are managed by the model creators directly rather than by the cloud platforms.[5]

Lastly, inference is also provided by companies specializing in AI compute services, such as Together AI, Cerebras, and Groq. These firms typically serve only open-source models and compete by optimizing software and hardware to enhance speed and reliability. For example, Cerebras claims to have "The Fastest AI Infrastructure".[6]

There are a variety of inference providers that serve the same model. These providers differ in pricing, context length, completion length, latency, and throughput. In addition, providers occasionally experience outages during which the service is unavailable, and users may consider uptime an important factor when choosing providers.

2.3 LLM API Pricing

Inference providers mainly meter usage along four primitives: prompt tokens (the prompt a firm sends), completion tokens (the text the model generates), reasoning tokens (additional internal deliberation steps in "reasoning mode"), and cache tokens (prompt tokens, such as system prompts, that are reused across API calls and are cheaper to serve). Providers quote price-per-million-token rates and add the metered components to determine the charge for a call. Cache pricing amortizes fixed prefix costs over repeated calls and turns a long prompt into a quasi-fixed prompt. Reasoning tokens are only present for models or modes that expose an explicit reasoning budget; standard decoding does not meter them separately.

Per token prices typically follow a clear hierarchy driven by marginal resource intensity.

[4] After the sample period covered in this paper, Azure began offering Anthropic models.

[5] For example, in addition to an $8 billion investment from Amazon, Anthropic announced AWS as its "primary training partner," which entails using AWS chips and compute resources to build, train, and deploy its models. See Tech Crunch.

[6] Cerebras holds the record for fastest AI inference using its computing power. As of September 2025, the company was valued at $8.1 billion. See Business Wire.

Reasoning tokens incur the highest unit cost because they require additional compute and memory beyond standard decoding. Completion tokens are next, as autoregressive generation is stepwise and scheduler-intensive. Prompt tokens are cheaper because encoding a prefix is a single pass. Cache writes are usually priced near the prompt price, reflecting a one-time encoding and storage operation. In contrast, cache reads are deeply discounted—often at the lowest price point—because they bypass recomputation and exploit locality.

Pricing also reflects competitive dynamics in the provision of inference for open-source models. Since open-source weights are freely available, multiple providers can host the same model and differentiate themselves on cost, latency, throughput, reliability, or value-added services. This often results in downward pricing pressure, with some providers offering substantially discounted rates or even free tiers for widely used open-source models. In contrast, closed-source models, in which only the creator controls access, tend to maintain higher, more stable prices.

2.4 LLM Aggregators and OpenRouter

LLM aggregators are platforms that sit between users and model providers, offering a single interface through which users can access many different models. Without an aggregator, users must integrate with multiple APIs separately, track heterogeneous pricing structures, and manage variability in latency, throughput, and uptime. Rather than requiring users to integrate separately with each provider's API, aggregators standardize access, often adding features such as usage analytics, routing across multiple providers, and pricing transparency. By lowering switching costs, aggregators play an important role in increasing competition among providers while also simplifying adoption for enterprises and individual developers.

OpenRouter is the leading LLM aggregator and describes itself as "the unified interface for LLMs," with "better prices, better uptime, no subscription." In practice, this means that OpenRouter provides a standardized API for invoking any of hundreds of models. This offers a range of advantages for developers, in addition to its simplicity. One advantage is that, for models offered by multiple providers (e.g., GPT-4o on Azure versus OpenAI), OpenRouter can dynamically route API calls based on latency, cost, or throughput. OpenRouter also allows developers to specify fallback models under defined conditions.

OpenRouter primarily derives revenue by charging a percentage fee on the value of each API call. This is done in one of several ways. If using an OpenRouter key, users must purchase credits and are charged 5.5% on the purchase price. If users bring their own key, for example, for the OpenAI API, they are charged 5% of what the API call would cost on OpenRouter. OpenRouter also partners with model developers to trial beta versions of models or to offer discounts.

On its website, OpenRouter provides influential rankings of the top models by token usage over time. Each model has a page that includes pertinent information such as price, uptime statistics, and top apps using the model per week (including tokens used). Apps and developers receive a 1% discount for allowing their API calls to be used for ranking.[7]

3 Data

In this section, we present an overview of the data.

3.1 OpenRouter Data

We obtained OpenRouter data through the Internet Archive and web scraping. The dataset can be divided into three main components.

The first component is model-level data, collected from each model's page on the OpenRouter website. This page reports, for each provider hosting the model, key features such as price, quality characteristics (latency and throughput), and additional attributes such as data retention policies. The model page also provides usage statistics, including total prompt and completion tokens over the last 90 days, the number of requests, and information on the model's creator and release date. In addition to total model usage, we also observe the number of tokens processed by each provider within a model for the top ten providers.

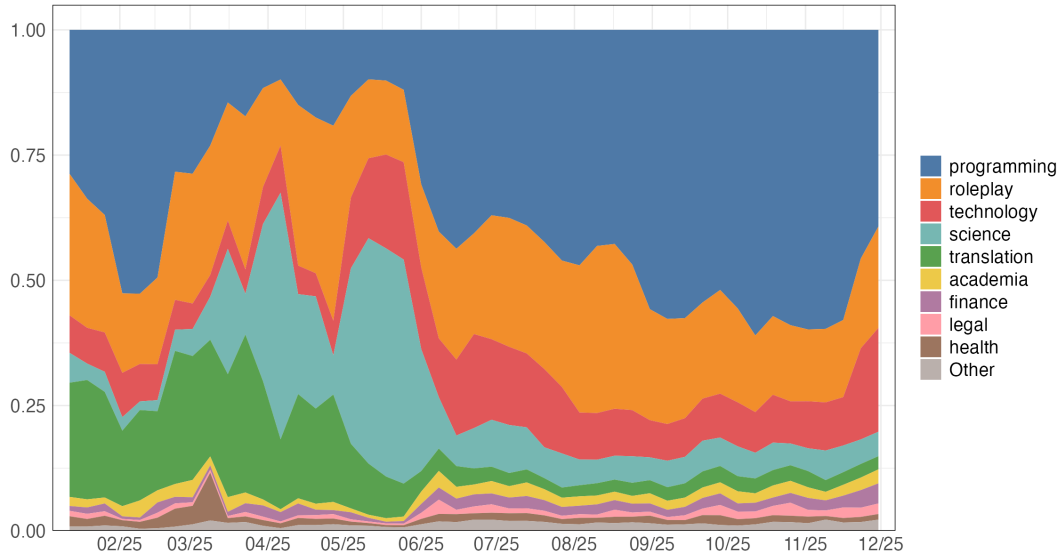
The second component is category-level data. OpenRouter samples and classifies a random sample of queries into use case categories, including legal, health, finance, academia, marketing, programming, trivia, translation, science, SEO, technology, and roleplay. For each category, the data report the total number of sampled API requests, the total number of prompt and completion tokens across these requests, and the distribution of those tokens across models.

The third component is user-level data. For each model, the platform publishes the top 20 users ranked by the number of tokens consumed. If a user appears among the top 20 users for any model, its usage across all models is reported on a separate page.

Our OpenRouter sample begins in July 2023. Between July 2023 and January 2025, the dataset contains gaps, after which a balanced daily panel becomes available. Because usage volumes are low before January 2025, all analyses of token usage restrict the sample to begin in January 2025. In contrast, the pricing and intelligence analyses use data starting in July 2023. Finally, pricing and token-usage information disaggregated at the provider level becomes available only after April 2025; before this point, we observe only the price of the advertised provider associated with each model on OpenRouter and each model's total token usage. We provide more information about data collection and processing in Appendix C.

[7] See https://openrouter.ai/docs/faq for Open Router's pricing and discount information.

Figure 1: Popularity of LLM Use Cases Over Time



Notes: This figure shows the weekly token share by use case category over time. Each line represents a different use case category, showing how usage patterns have evolved across different application domains. The x-axis is calendar time, and the y-axis reports the share of total tokens.

Figure 1 shows the share of usage categories in terms of total tokens used over time. The top category is programming, accounting for nearly 50% of recent usage. The second- and third-largest categories are roleplay and technology, each accounting for roughly 15% of total processed tokens. The remaining categories have substantially smaller shares. This composition of use cases likely does not reflect the overall distribution of AI use in the broader economy; our sample is disproportionately weighted toward technology firms and programming applications. This is important to keep in mind when interpreting some of our results.

Lastly, there are three nuances worth noting about the OpenRouter data. First, some companies trial masked versions of their models prior to an official release. For example, "Optimus Alpha" was actually OpenAI's GPT-4.1. These requests are included in our token counts but are treated as separate models throughout our analysis. Second, providers offer free access to certain models, often with rate or capacity limits. Where appropriate, we exclude free requests from our analysis, but note that there may be some measurement error. Third, model names in the raw data required substantial manual cleaning and aggregation (See Table OA-7 for examples). We discuss this cleaning procedure and other data processing steps in Appendix D.

3.2 Microsoft Azure Data

The Microsoft Azure data come from its AI Foundry service, which is Azure's commercial offering that allows customers to access individual models through an API in a "model-

as-a-service" paradigm. Unlike OpenRouter, which routes users to LLMs hosted by third-party providers, Microsoft directly hosts these models, and not all models are available on the platform. The primary closed-source models are those developed by OpenAI and xAI, whereas the open-source offerings include models from Meta, DeepSeek, and other open-source providers. Publicly available sources indicate that Azure AI Foundry processed 100 trillion tokens in the second quarter of 2025.[8]

The Azure dataset records, at the firm-day level, usage across several dimensions: the number of prompts, completions, cache, and, where available, reasoning tokens consumed for each model. In addition, it contains information on customer industry classifications and model prices. The data are available from July 2023 through June 2025, enabling analysis of LLM usage patterns in a long panel that begins near the commercial introduction of large language models. A limitation of this dataset is that it reflects only a restricted set of models used by a selected group of Azure customers.

3.3 Pricing Data

We collect pricing information from OpenRouter. OpenRouter reports the price of each model separately for prompt and completion tokens on its model page. These prices are identical to those charged by the inference providers, as OpenRouter retrieves them directly from the providers' APIs.[9] This allows us to construct a pricing dataset covering several hundred models over the past two years, constructing the most comprehensive data on prices to the best of our knowledge. The pricing data start in July 2023 and are available with some gaps until November 2024. From this point onward, we observe prices daily.

Some limitations of the pricing data are worth noting. First, OpenRouter has only recently begun reporting cache pricing, which has become increasingly common. As a result, we do not have historical cache prices across models. Second, for certain providers—particularly cloud platforms—prices may vary across regions. In these cases, we observe only the price from the region selected by OpenRouter.

3.4 Benchmark Data

Benchmarks are commonly used to evaluate model capabilities, with different benchmarks targeting distinct dimensions, such as coding, mathematics, and general reasoning. While they have been highly useful in standardizing evaluation and enabling systematic comparisons across models, many existing benchmarks are beginning to show signs of saturation, making it harder to distinguish incremental improvements. In addition, there are challenges, such

[8]Source: Azure Blog Post.
[9]See OpenRouter Provider Documentation.

as potential arbitrariness in design and the possibility of question spillover, whereby models may have prior exposure to related data.

Nevertheless, models differ substantially in their capabilities, and it is helpful to quantify these differences systematically. To avoid relying on any single arbitrary benchmark, we draw on benchmark indices from Artificial Analysis, a widely used benchmark provider. Artificial Analysis reports results across 13 benchmark measures—including *MMLU Pro*, *GPQA*, *HLE*, *LiveCodeBench*, and *SciCode*—and aggregates eight of them into a composite measure called the *Artificial Analysis Intelligence Index* (hereafter, the Intelligence Index).[10] While the Intelligence Index captures overall intelligence rather than distinct dimensions, the underlying benchmark indices are highly correlated (see Figure OA-2), suggesting that the choice of benchmark is unlikely to materially affect our results. We match this metric to both of our datasets, achieving coverage of 48.9% of models and 87.2% of tokens in the OpenRouter data.

3.5 Discussion and Limitations of Datasets

While we believe our paper draws on the most comprehensive data available on firms' use of LLMs, it is important to acknowledge and discuss the limitations of our data sources. The main strength of the OpenRouter data is its breadth of model coverage, as it provides access to nearly every model on the market for API use. This enables us to combine usage data across models from different providers—something that would otherwise not be possible.

The main limitation of this dataset is that usage data are available primarily at the aggregate level. Another important limitation concerns the selection of users across platforms. OpenRouter's primary customer base comprises app developers who create AI-based applications for mobile platforms and websites, resulting in a user base that is disproportionately young and startup-oriented. We mitigate these limitations by using firm-level usage data from Microsoft Azure, which reflects a different composition of firms, likely including more established, enterprise-level customers.

A word of caution is also warranted when interpreting the Intelligence Index and other benchmarks. There is no guarantee that such indices are cardinally scaled in an economically meaningful way. In particular, a one-unit increase from 20 to 21 may represent a different improvement in difficulty or utility than an increase from 60 to 61.

4 Supply: Models, Creators, and Inference Providers

This section examines the supply side of LLMs. We begin by analyzing the number of models, creators, and providers over time, and then turn to changes in model capabilities. Our goal is

[10] See OA-3 for an overview of each of the 13 benchmarks and which ones are included in the Intelligence Index.

to document the supply of models that are readily accessible via API—rather than all trained models or wrapper variants. Because OpenRouter captures nearly the whole universe of such API-accessible models, the models listed on OpenRouter provide a useful measure for this purpose.

4.1 Models and Creators

Figure 2(a) shows the cumulative number of models over time that were available in OpenRouter, separated into closed- and open-source categories. Since the introduction of GPT-3.5 in late 2022, the number of models has grown exponentially, exceeding 600 by late-2025. Much of this growth occurred in the last year and a half, with the total rising from about 60 in early 2024 to more than 650 by December 2025. The expansion is driven primarily by open-source models, which proliferated after July 2023 and quickly surpassed closed-source models. As of December 2025, there are 434 open-source models compared to 217 closed-source models ever observed in OpenRouter.[11]

Figure 2(b) reports the cumulative number of model creators, again separated by open- and closed-source. As of December 2025, there are 85 unique creators in the dataset, the majority of which are open source. The number of open-source creators increased sharply after July 2023 and quickly exceeded that of closed-source creators. By December 2025, there are 75 open-source creators compared to 10 closed-source creators, suggesting that the surge in open-source models reflects not just a few prolific creators but also a broader dispersion of contributors entering the ecosystem.[12] It is also worth noting that the number of closed-source model creators remained at 10 since April 2025, indicating a maturing market where new closed-source entrants appear to face diminishing returns.

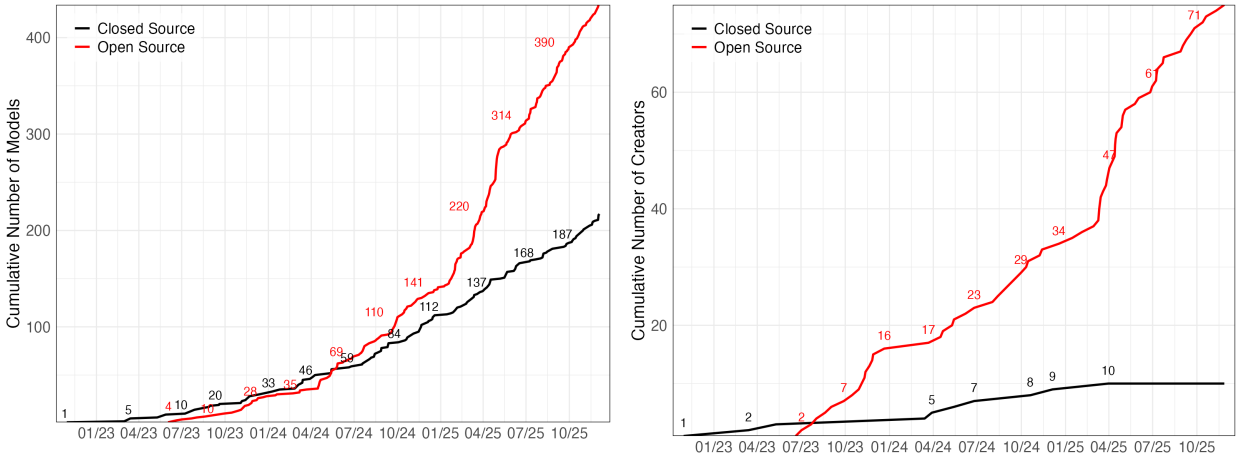
In Appendix Figure OA-3, we report the cumulative number of models created over time by different creators. The figure highlights the rapid scaling of several leading model creators. Early growth was driven by OpenAI, Anthropic, and Mistral, but from mid-2024 onward, we observe a sharp increase in contributions from other providers, such as Google, Qwen, and DeepSeek. By mid-2025, Qwen and Google surpass all other creators in the number of released models, each exceeding 50.[13]

[11] Some of these models have been deprecated over time and are no longer available, though the deprecation rate is lower than one might expect.

[12] The number of open-source model creators is, of course, a significant underestimate, as there are hundreds of thousands of open-source models from thousands of creators on Hugging Face. Our focus here is on models that are ready to use via an API, a capability more relevant to enterprise AI adoption.

[13] Figure OA-4 reports the availability of model families over time, and Figure OA-6 reports the availability of individual models over time for the top 10 creators.

Figure 2: Cumulative Growth of LLMs and Creators



(a) Number of Models Created, by Type **(b)** Number of Creators, by Model Type

Notes: Subfigure 2(a) shows the cumulative number of models and Subfigure 2(b) shows the cumulative number of creators, labeled for every 3 months from October 2022 through December 2025

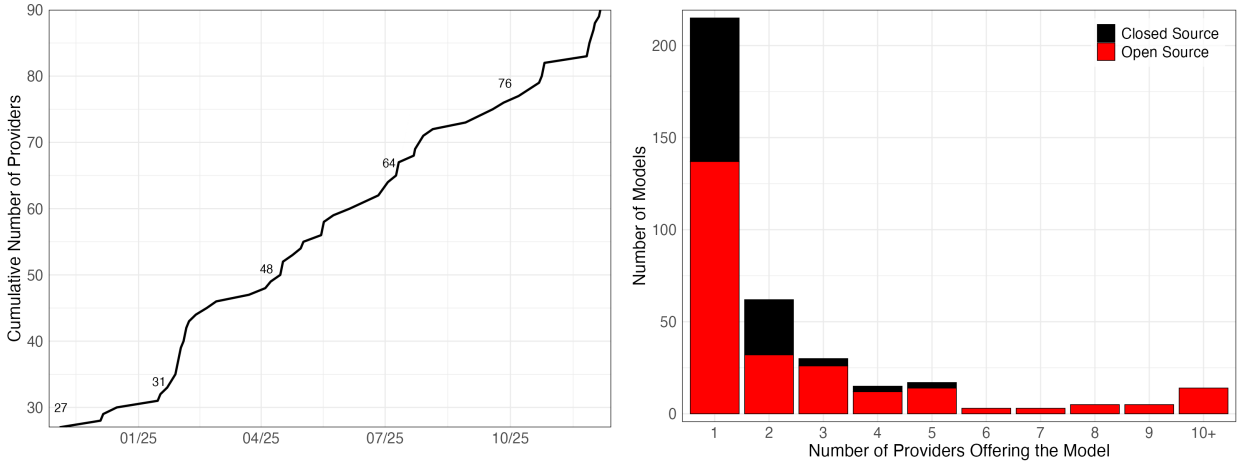
4.2 Inference Providers

We next turn to inference providers. As described in Section 2, LLM inference can be offered by multiple providers beyond the original model creator. The left panel of Figure 3 shows the cumulative number of inference providers ever observed in the OpenRouter data. In November 2024, there were 27 providers. This number increased steadily through 2025, reaching 39 by February, 48 by April, and 90 by December. The rapid growth in providers is driven primarily by entrants offering inference for open-source models.

How does the entry of inference providers affect competition in model provision? Figure 3(b) reports the distribution of models by the number of providers offering them, separately for open- and closed-source models. Most closed-source models are offered by a single provider, typically the original creator—for example, Gemini models are available only through Google. Thirty closed-source models are distributed by two providers; most are OpenAI models offered on both OpenAI's platform and Microsoft Azure. A smaller set of models is available through three to five providers. These include Anthropic's models, hosted by third-party providers such as Google, Amazon, and Microsoft, and xAI's models, which are accessible via Azure and Amazon, as well as via xAI itself.

By contrast, open-source models are far more widely distributed: 137 are hosted by a single provider, 32 by two providers, and several are hosted by three to ten or more providers. Notably, 14 open-source models are available from at least ten providers. This reflects the fundamental difference in competition between closed- and open-source models. Closed-source models are available only through contractual arrangements with their creators. In

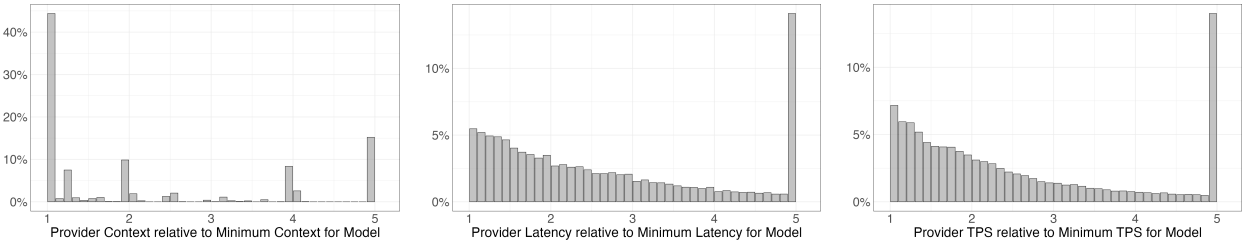
Figure 3: Cumulative Growth of Inference Providers and Multi-Hosted Models



(a) Number of Inference Providers **(b)** Number of Providers per Model, by Type

Notes: Subfigure 3(a) shows the cumulative number of providers, labeled every 3 months from November 2024 through December 2025, and Subfigure 3(b) shows provider distribution by model type for the last 30 days (November 7, 2025 through December 6, 2025).

Figure 4: Distribution of Provider Performance Metrics for Open-Source Models



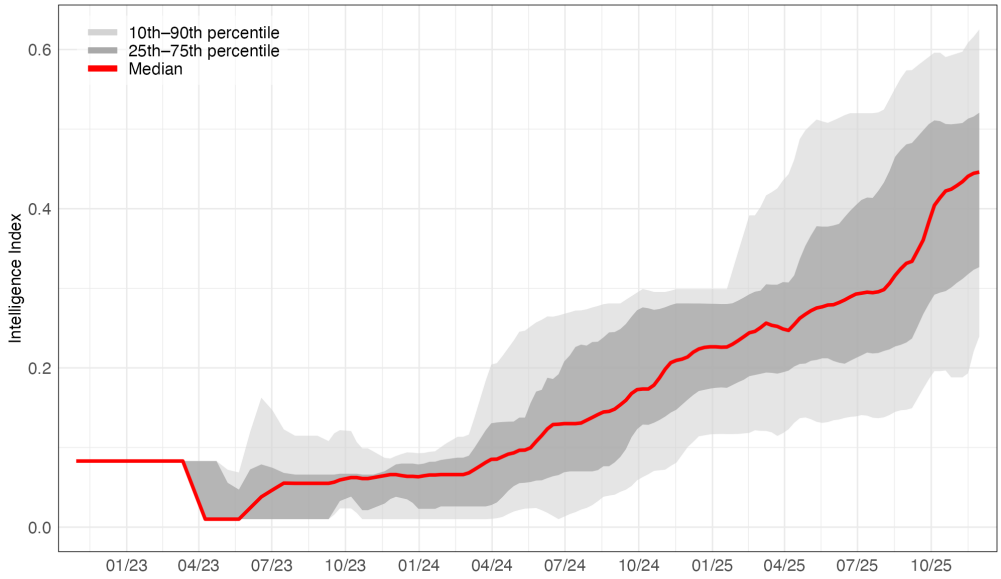
(a) Context Window Ratios **(b)** Latency Ratios **(c)** Throughput Ratios

Notes: Distributions include open-source models with more than two providers, and ratios are winsorized at 5. The ratio is normalized relative to the value of the minimum provider offering the same model on that dataset. The distribution is at the date–provider level and covers March 2025 to December 2025.

contrast, any firm with sufficient GPU capacity can download the weights of an open-source model and offer inference services. These results suggest that entry rapidly leads to multiple providers offering identical open-source LLMs.

While providers of open-source models offer the same underlying weights, they can differ across many dimensions—primarily price (which we analyze in the next section), quality metrics such as latency, and other characteristics, including the provider's nationality and its data-use policies. Figure 4 examines differentiation across these quality metrics. For each model-provider-date, we calculate a performance ratio by dividing each provider's value by the minimum value for that model-date, thereby measuring performance relative to the best performer (for latency) or the worst performer (for throughput and context window). We then plot the distribution of these ratios across all model-provider-dates.

Figure 5: The Improving Intelligence of New Models



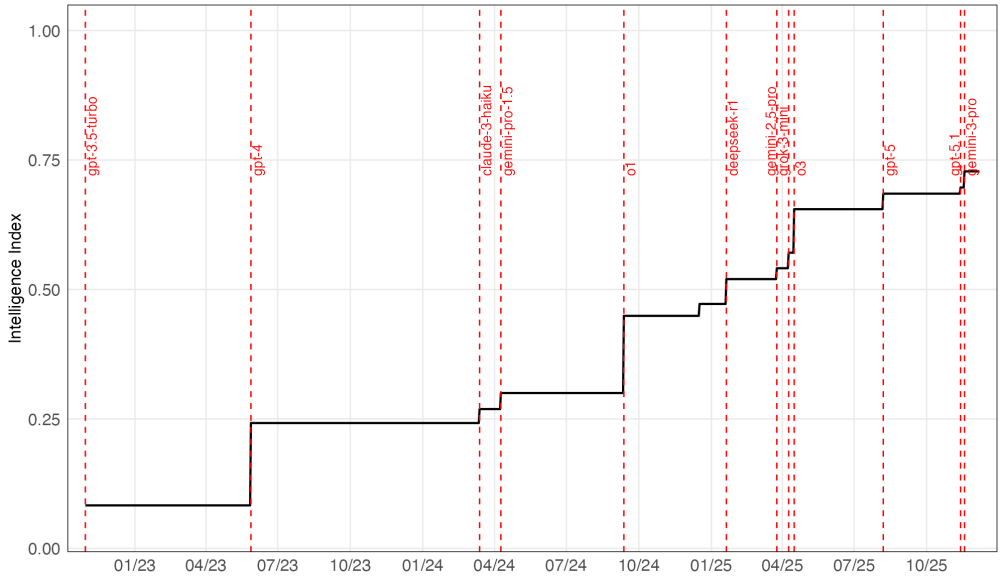
Notes: This figure shows the distribution of Artificial Analysis's Intelligence Index for models created within 6 months of each plotted date, starting from October 2022. The red line is the median, the dark shaded area is the 25th-75th percentile range, and the light shaded area is the 10th-90th percentile range.

The plots reveal substantial differentiation across providers in latency and throughput, and comparatively less in context window. For throughput, around 29% of providers are within 50% of the lowest-throughput provider, while 14% achieve throughput more than five times higher. Differentiation is similar for latency: about 25% of providers fall within 50% of the lowest-latency provider, while 14% exhibit latencies more than five times higher. By contrast, context window sizes show less variation, with approximately 44% of providers offering identical context windows for a given model. These results suggest that price is not the only factor differentiating open-source model providers; other technical capabilities also play an important role in shaping competition among providers. In the next section, when we analyze pricing, we examine how these characteristics influence demand.

4.3 Changes in Model Capabilities

We next analyze how the capabilities of newly launched models have changed over time. Figure 5 shows the distribution of model capabilities for models released within the past six months of the date shown on the x-axis, based on the Intelligence Index. The red line reports the median, while the dark and light gray shaded areas indicate the 25th–75th and 10th–90th percentile ranges, respectively. The results suggest a clear upward trend in the median model Intelligence Index over time, with the median capability increasing from approximately 0.1 when GPT-3.5 was launched to over 0.4 today. The widening distribution further reflects gains at the frontier: today's top-performing models achieve roughly six times

Figure 6: Most Intelligent Models Over Time



Notes: This figure displays the maximum of Artificial Analysis's Intelligence Index each day from October 2022 through December 2025. The vertical lines show when a new model becomes the top performer.

the performance of the earliest models in the sample, and there are now multiple models at this level. At the lower end of the distribution, however, we continue to observe the launch of relatively weaker models. This indicates that, even as overall performance improves, the supply of models is diversifying to meet different capability requirements.

Figure 6 plots the performance of the top model over time, with the frontier model annotated by name. In the early period, GPT-3.5 Turbo and GPT-4 each maintain state-of-the-art performance for roughly a year, with index scores of 0.08 and 0.24, respectively, illustrating OpenAI's early leadership. Beginning around April 2024, other providers briefly take the lead—first Claude 3.5, followed shortly by Gemini 1.5 Pro—both representing incremental gains.

A subsequent phase is marked by the emergence of "reasoning" models. The introduction of o1 produces a discrete jump in the index, comparable in magnitude to the step from GPT-3.5 Turbo to GPT-4. Several additional reasoning models follow, delivering further—though smaller—improvements. Overall, Figure 6 highlights OpenAI's early dominance, the rapid catch-up by competitors, and, more recently, a pattern of more frequent leadership changes accompanied by incremental advances at the frontier. This suggests both the industry's dynamism and the presence of several firms operating at high levels of intelligence.

5 Pricing

In this section, we document pricing patterns in the market for LLMs and examine how prices have evolved over time both within and across models. We show that substantial and persistent price differences exist even among models with similar benchmark performance. A key explanatory factor is whether a model is open- or closed-source, with open-source models priced substantially lower per unit of measured intelligence. Lastly, we analyze market entry and the resulting price competition among providers that host open-source models.

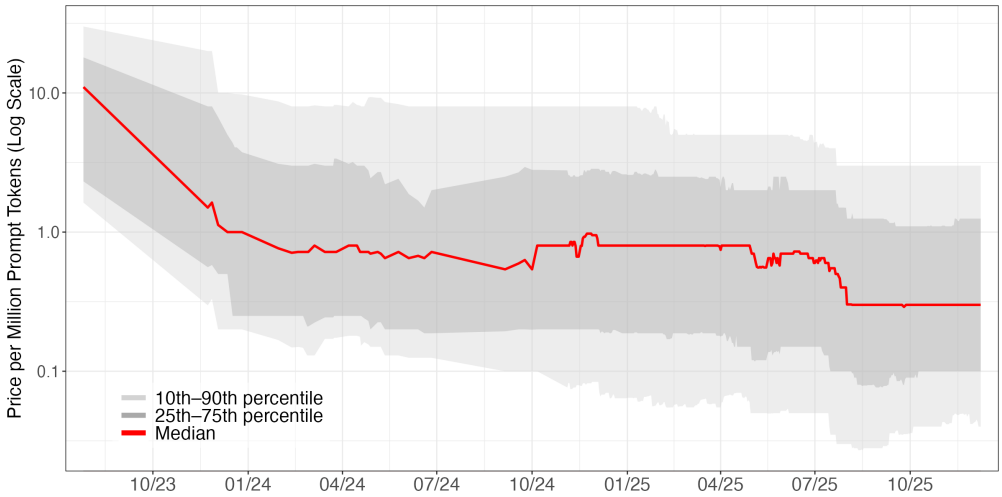
5.1 Evolution of Pricing

Figure 7 shows the distribution of prompt and completion prices over time. Following a sharp decline between mid-2023 and early 2024, the cross-sectional distribution of prices has remained relatively stable even as newer, more capable models entered the market. This pattern suggests that quality-adjusted prices have continued to decline, a point we analyze in greater detail below. The figure also reveals substantial price dispersion across models: at any given point in time, models in the bottom decile are between 50 and 150 times less expensive than those in the top decile. Such heterogeneity reflects not only vertical differentiation in model quality but also other factors, including whether a model is open- or closed-source and inference-related attributes such as latency and throughput.

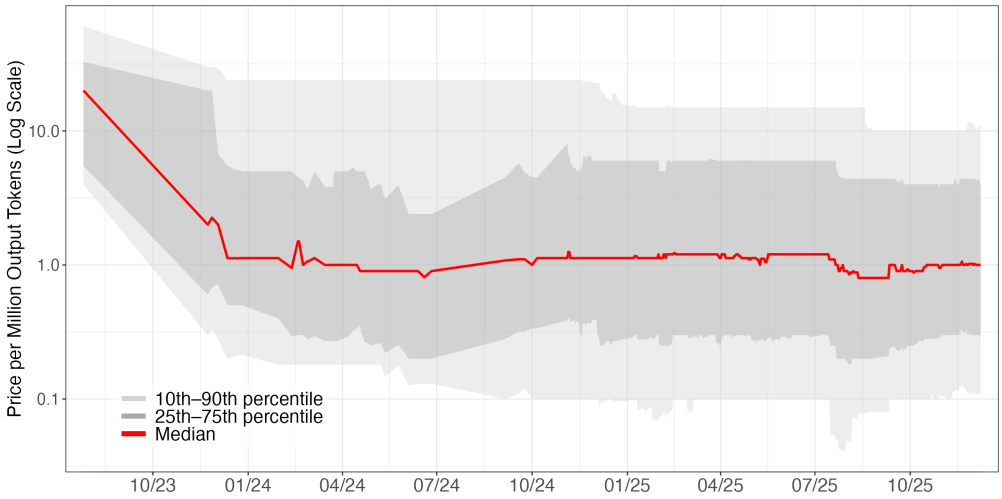
Figures OA-8 to OA-10 in the Appendix plot the price trajectories of Anthropic, Google, and OpenAI models over time. A clear pattern emerges: closed-source models exhibit stable pricing, with most variants maintaining their launch price throughout their lifecycle. When new versions are released, they often enter at prices comparable to those of their predecessors. For example, Claude 3.7 Sonnet and Claude 4 Sonnet were introduced at nearly identical prices, consistent with Anthropic's broader pricing strategy. By contrast, open-source models show considerably more volatility, with frequent price adjustments and downward revisions over time.

This pattern is further confirmed in Figure 8(a), which reports the probability that a model experiences a price change within a week by model type. Across all models, open-source models change prices in 5.5% of provider-model-week observations, more than double the rate of closed-source models at 1.8%. When weighting by usage, the difference is even more stark: open-source token prices change in 10.5% of provider-model-token-week observations, compared to 0.9% for closed-source models. Together, these figures highlight a systematic difference in pricing dynamics: closed-source providers maintain rigid, stable pricing across model generations, whereas open-source models are subject to active competition among

Figure 7: Distribution of Token Pricing Over Time



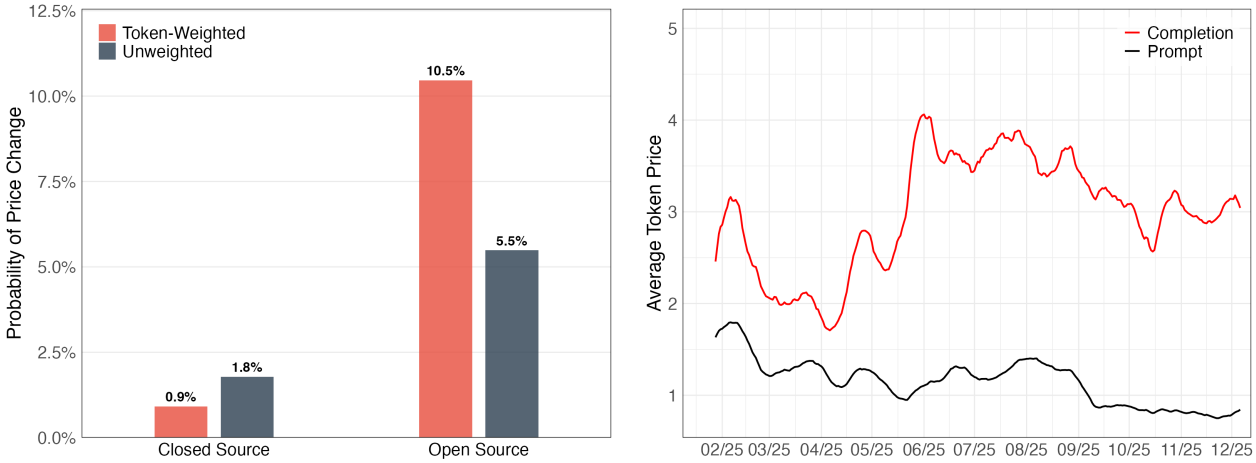
(a) Prompt Token Pricing



(b) Completion Token Pricing

Notes: Distribution of (unweighted) token prices over time on a log scale. The sample includes models available on OpenRouter on the date shown on the x-axis and excludes free models. Subfigure 7(a) plots prompt token prices per million tokens; Subfigure 7(b) plots completion token prices per million tokens. The red line shows the median, the dark shaded band shows the 25th–75th percentile range, and the light shaded band shows the 10th–90th percentile range.

Figure 8: Probability of Price Change and Token-Weighted Prices Over Time



(a) Probability of Weekly Price Change by Model Type

(b) Token-Weighted Average Prompt and Completion Prices Over Time

Notes: Subfigure 8(a) shows the probability that a provider-model combination changes price within a week, reported separately for open-source and closed-source models. Both unweighted and token-weighted probabilities are displayed. Subfigure 8(b) reports token-weighted average prompt and completion prices for 2025, computed using 14-day rolling window averages. Completion prices are weighted by completion tokens, whereas prompt prices are weighted by prompt tokens.

providers, leading to greater price fluctuations and lower average prices.[14]

Although the price distribution across models remains relatively stable, the effective prices developers pay may differ when usage patterns are accounted for. To measure this, we compute a token-weighted average price.[15] Figure 8(b) shows the resulting series for 2025. Completion prices exhibit noticeable cyclical variation, while prompt prices are consistently lower and more stable. On average, prompt token prices remain close to $1, whereas completion token prices fluctuate between about $2 and $4 over the year.

5.2 Price of Intelligence

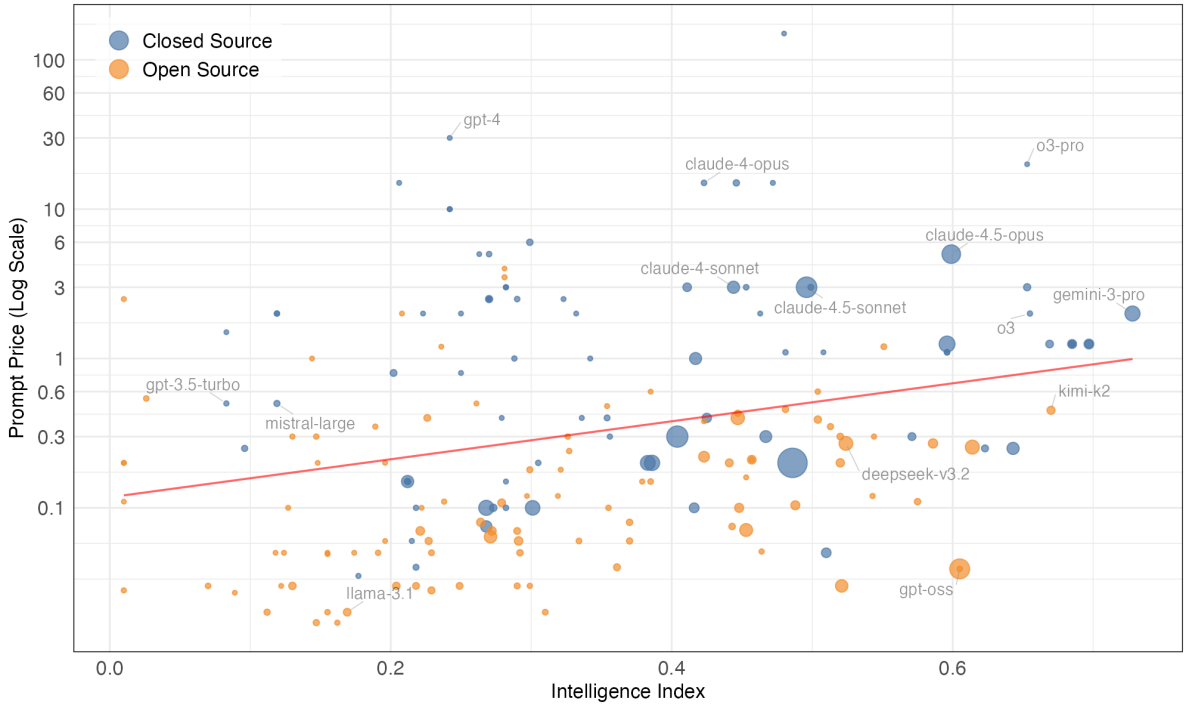
In the previous section, we showed that average prices paid are fairly stable over time. That said, the intelligence of models has increased drastically over our sample period. In this section, we document trends in the price of intelligence.

As previously mentioned, Artificial Analysis generates indices of model capabilities by aggregating results across a variety of benchmarks. The Intelligence Index incorporates performance across eight benchmarks: MMLU-Pro (massive multitask language understanding),

[14] See Table OA-1 for regression results supporting these findings. Figure OA-7 plots the price trajectories of the top 5 most popular open (dashed) and closed-source (solid) models on OpenRouter. The prices of closed-source models remain constant throughout their lifetimes, and new models enter at prices similar to those of previous models. For example, Claude 3.7 Sonnet and Claude 4 Sonnet have identical pricing. In contrast, open-source models exhibit substantial price fluctuations, with negative price trajectories on average.

[15] To calculate this number, we sum price times token quantity and divide by the total number of tokens. We do not observe cached tokens from OpenRouter, so any caching discounts are excluded from this calculation.

Figure 9: Prompt Price vs. Intelligence



Notes: Scatter plot of prompt price (log scale) against the AI Intelligence Index. Point size is proportional to total usage; colors indicate closed- vs. open-source models. The red line shows a fitted linear trend. The date of this snapshot is December, 6, 2025.

GPQA Diamond (graduate-level scientific reasoning), Humanity's Last Exam (comprehensive academic assessment), LiveCodeBench (real-time coding evaluation), SciCode (scientific programming), AIME (mathematical problem-solving), IFBench (instruction following), and AA-LCR (long-context logical and commonsense reasoning). These benchmarks collectively assess models' capabilities across diverse domains, from mathematical reasoning and scientific knowledge to coding proficiency and instruction comprehension. The Coding Index is based solely on LiveCodeBench and SciCode. Figure OA-1 illustrates the trends in the Intelligence Index and Coding Index of tokens generated in 2025. We see a steady increase in the intelligence and coding ability of tokens generated in this period.

Figure 9 displays a scatter plot of price versus intelligence per model as of December 2025, with the y-axis shown on a logarithmic scale. Each point corresponds to a model, where color denotes whether the model is open- or closed-source, and point size reflects total usage.

Two patterns emerge. First, the relationship between cost and intelligence is linear in logs, implying a nonlinear relationship in levels: incremental improvements in intelligence are associated with disproportionately higher costs at the upper end of the distribution. Second, there is substantial price dispersion conditional on intelligence. For nearly any given intelligence score, prices vary by up to two orders of magnitude, and several high-priced

Table 1: Price-Intelligence Regression Results

	Log Price per Million Prompt Tokens				
	(1)	(2)	(3)	(4)	(5)
Intelligence Index	0.039***	0.029***	0.040***	0.052***	0.047***
	(0.009)	(0.010)	(0.011)	(0.014)	(0.017)
Open Source		-2.46***	-2.17***	-2.00***	-1.91***
		(0.266)	(0.267)	(0.458)	(0.440)
Supports Reasoning		-0.107	0.132	-0.146	0.097
		(0.307)	(0.315)	(0.367)	(0.371)
Log Context Length		-0.347**	-0.260**	-0.273*	-0.323**
		(0.134)	(0.130)	(0.154)	(0.149)
Intelligence Index × Model Age 120–360 Days					-0.004
					(0.019)
Intelligence Index × Model Age 360+ Days					0.085***
					(0.025)
R^2	0.101	0.420	0.457	0.587	0.615
Observations	152	152	152	152	152
Model Age Bin fixed effects			✓	✓	✓
Creator fixed effects				✓	✓

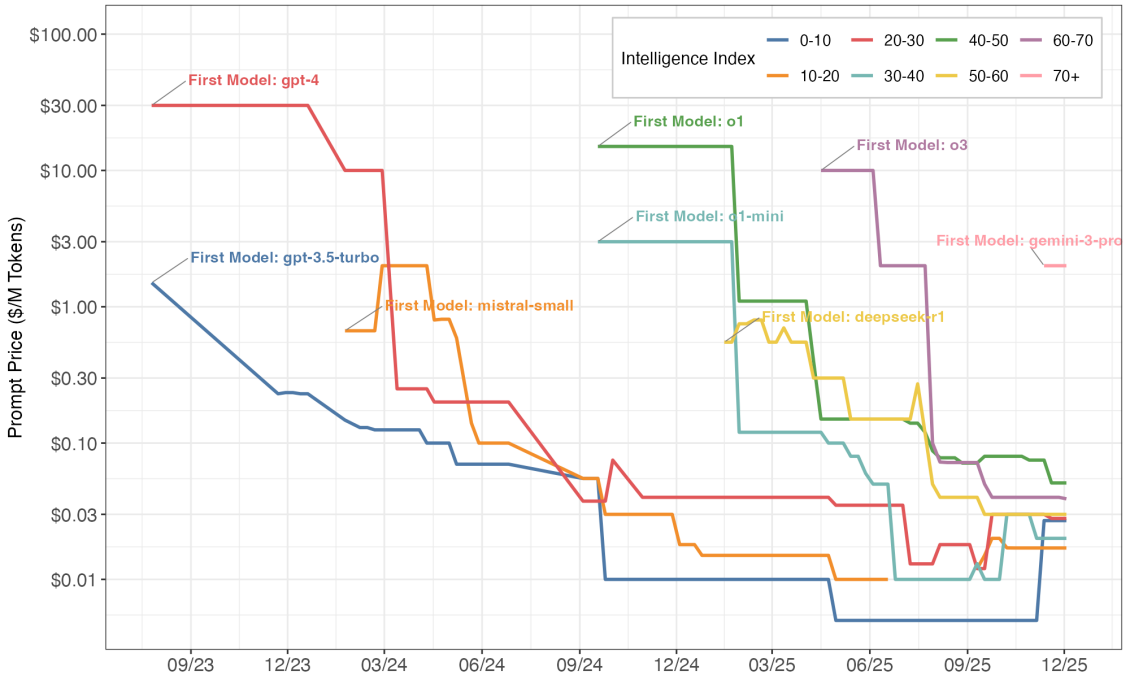
Notes: This table reports regression results examining the relationship between model prices and intelligence scores. The dependent variable is the log of prompt token price (per million tokens). The main independent variable is the Intelligence Index from Artificial Analysis. Controls include model characteristics (open source indicator, context window size), temporal effects (days since model creation), and creator fixed effects. Standard errors clustered at the model level are reported in parentheses.

models remain in use. An important determinant of this dispersion is whether a model is open- or closed-source. Closed-source models systematically lie above the fitted trend line, suggesting that they command a significant premium relative to open-source models with comparable intelligence levels.

To investigate these patterns further, we conduct regression analyses of price on intelligence and other factors. Table 1 presents the results of regressing log price per million prompt tokens on intelligence and other model characteristics. Column (1) shows a simple regression of price on the Intelligence Index, while the remaining columns add controls for open-source status, reasoning capabilities, context length, and creator fixed effects, as well as interaction terms.

The regression results confirm that more intelligent models command higher prices: a one-point increase in the Intelligence Index is associated with a 3.9% increase in price. In column (2), covariates for the model's open-source status, reasoning capabilities, and context length account for 32% of the variation in prices. Adding creator and model-age fixed effects further increases the explanatory power, and the full specification explains roughly 61% of the variation in prices.

Figure 10: Minimum Price of Models of a Given Level of Intelligence



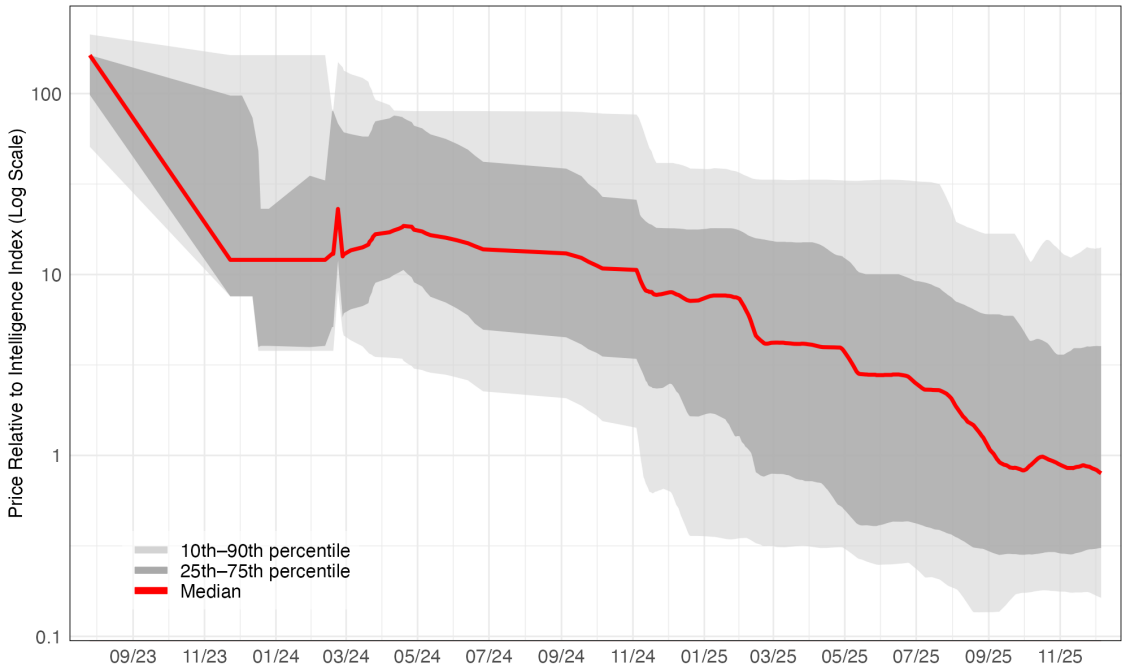
Notes: This figure shows the minimum price for models at each level of intelligence over time. Each line represents a distinct bin of intelligence scores. The y-axis uses a log scale to display prompt price per million tokens, and the x-axis shows calendar time.

Another important finding from Table 1 concerns the pricing of open-source models. Conditional on intelligence levels, open-source models are 87% cheaper than closed-source models ($1 - \exp(-2)$). As shown in Figure 9, despite these lower prices, the market share of open-source models remains substantially below that of closed-source alternatives. There are two potential explanations for this gap. First, the Intelligence Index may not fully capture important differences in model capabilities. Second, users may systematically discount the value of open-source models—either because of actual or perceived differences such as weaker brand reputation, privacy concerns, or less extensive customer support.

Next, in Figure 10, we plot the minimum price for a given level of intelligence and how it has evolved over time. We observe steep declines in pricing across all levels of intelligence. Perhaps most impressively, the price for a GPT-4 class model has fallen by a factor of 1000 over the course of 2 years. For more recent state-of-the-art reasoning models, such as OpenAI's o1, the price drops are, if anything, more rapid.

Not all models released in a given time period are state-of-the-art in terms of intelligence. In Figure 11, we plot the distribution of the price-to-intelligence ratio for models released within 6 months of each date. For this figure, we are implicitly assuming that unit increases in the Intelligence Index are comparable to each other. Although this is surely not exactly

Figure 11: Distribution of Price-to-Intelligence Ratio Over Time



Notes: The figure reports the distribution of the price-to-intelligence ratio (14-day rolling average), constructed from all models available on OpenRouter on the corresponding date on the x-axis. The red line represents the median, the dark shaded area the 25th–75th percentile, and the light shaded area the 10th–90th percentile.

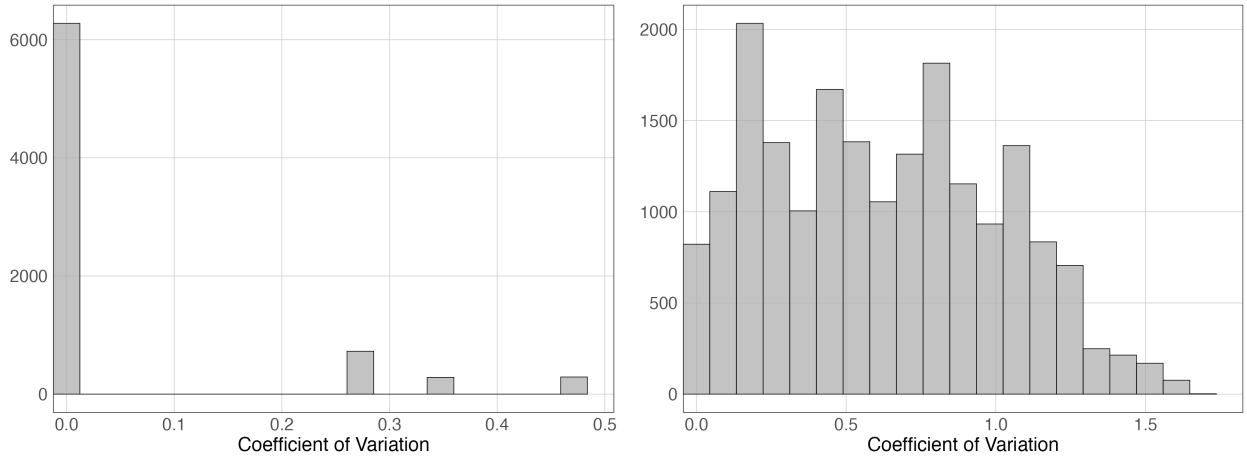
true, it allows us to compare the price distribution of models over time. As in the prior figure, this figure shows a dramatic decline in the cost of intelligence over time. The figure also reveals two distinct downward phases. The first occurs with the release of GPT-4, which was priced lower than GPT-3.5 despite achieving over twice the intelligence score. After a period of relative stability through much of 2024, prices resume their steady decline with the introduction of reasoning models.

Figure 11 also shows that prices are increasingly heterogeneous over time. As of December 2025, the price per unit of intelligence ranges from 0.01 to 20. This dispersion reflects at least two important factors. First, the relationship between price and intelligence is nonlinear: the cost of additional intelligence tends to increase with model capability. Second, providers and users may value dimensions other than price, such as latency, throughput, or model accessibility. We analyze the pricing–intelligence relationship in the remainder of this section and turn to these additional factors in the next section.

5.3 Pricing of Open-Source Models and Variation Across Providers

So far, we have analyzed overall pricing trends across different LLM models. Another important source of variation lies within the models offered by multiple providers. As discussed in the previous section, many models—particularly open-source models—are hosted by multiple providers. Understanding how competition among these providers translates into pricing

Figure 12: Coefficient of Variation of Same Model Prices Across Providers



(a) Closed-Source Models **(b)** Open-Source Models

Notes: Histograms of the coefficient of variation (CV = standard deviation divided by mean) of token prices across providers for the same model. Subfigure 12(a) shows closed-source models; Subfigure 12(b) shows open-source models. The x-axis is the CV of provider prices; the y-axis is the count of model observations. Included models are those with more than one provider on a given day. The sample period is from March 2025 to December 2025. The unit of observation is model-date.

differences is, therefore, an important question.

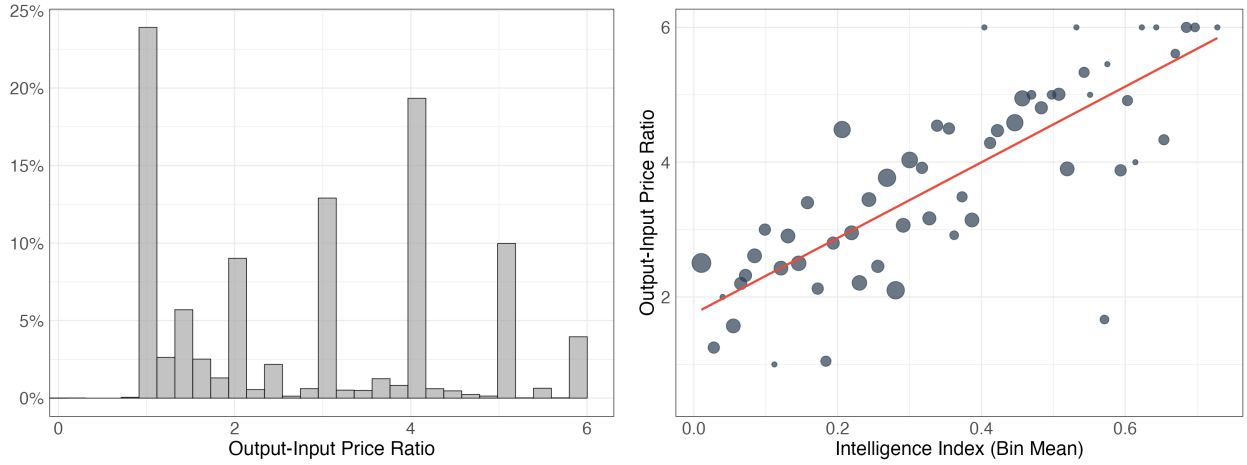
Figure 12 reports the distribution of price variation within models across providers, separately for closed- and open-source models. For closed-source models (Figure 12(a)), we observe very little variation: in most cases, the coefficient of variation is close to zero. This suggests that alternative providers price these models nearly identically to the creator. This result is unsurprising, as closed-source models are typically distributed to third parties under contractual agreements that stipulate uniform pricing (with the notable exception of Microsoft, which has unique access to OpenAI models)[16].

In contrast, Figure 12(b) shows substantial variation in prices across providers for open-source models, with coefficients of variation reaching as high as 1.7 for some cases. This finding is striking, given that the underlying models are technically identical across providers. While some differences in inference quality, such as latency, throughput, or reliability, may partly explain this variation, the magnitude of the observed price dispersion is unexpectedly large. In Appendix Table OA-6, we examine the correlation between prices and these attributes. We find that price is positively and statistically significantly related to throughput, whereas the correlations with the other attributes are not statistically significant.

[16]For OpenAI models served by Microsoft, we observe that the prices match those charged directly by OpenAI for the same model.

[16]There is scope for providers to serve slightly different versions of a model, by limiting the context window or changing the floating point precision.

Figure 13: Completion-to-Prompt Price Ratios and their Relationship to Intelligence



(a) Distribution of Price Ratios

(b) Price Ratio vs. Intelligence Index

Notes: Subfigure 13(a) shows the distribution of completion-to-prompt price ratios across providers (winsorized at 6). The unit of observation is model-provider-date. Subfigure 13(b) bins models by intelligence score and plots the ratio averaged across model-provider-date within each bin, with point size proportional to the number of models in the bin.

5.4 Pricing of Prompt and Completion Tokens

Another useful measure for understanding LLM pricing is the ratio of completion (output) to prompt (input) prices. Because the relative intensity of prompt versus completion usage varies substantially across categories, this ratio is an important determinant of the effective price users face. Figure 13(a) shows the distribution of this ratio across models. There is substantial heterogeneity: while many models have roughly equal prompt and completion prices, a sizable share exhibit higher completion costs. The median completion-to-prompt price ratio is 3, and 50% of models have ratios between 2 and 5.

Figure 13(b) illuminates this variation by showing the relationship between the completion-to-prompt price ratio and Intelligence Index. A clear positive relationship emerges: more capable models exhibit larger price differentials, with completion tokens commanding a higher premium relative to prompt tokens. Providers therefore systematically differentiate pricing by model capability, charging relatively more for completion tokens in higher-intelligence models.

6 Market Dynamics and Differentiation

This section analyzes the market dynamics for LLM models. Because our data does not cover the entire market, the reported market shares and token usage should not be interpreted as representative of the global market. Therefore, instead of focusing on precise market shares, we use observed patterns to draw inferences about the industry's competitiveness and firms'

model choices, with the goal of understanding potential vertical and horizontal differentiation.

With this caveat in mind, we focus on three types of analyses: (i) changes in dominant models and market shares over time, (ii) concentration of models, and (iii) the tendency of users to multi-home across models. Overall, the results indicate strong vertical and horizontal differentiation among models, as well as a highly dynamic industry that evolves rapidly with new model launches.

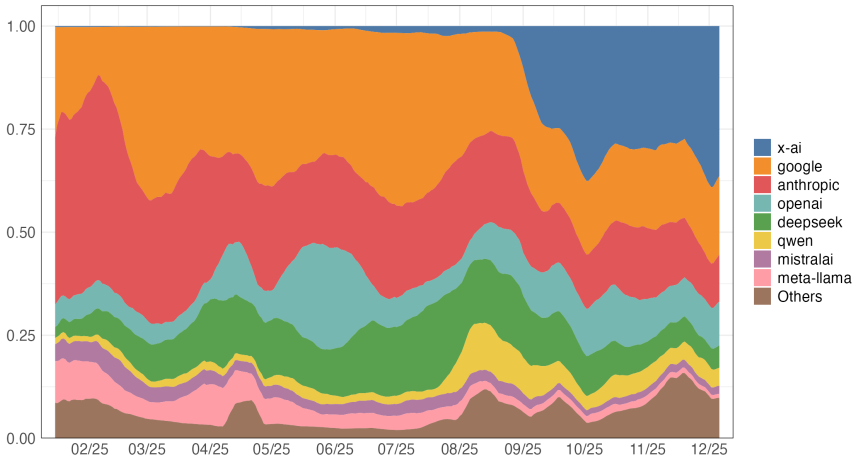
6.1 Market Shares over Time

Figure 14 illustrates aggregate usage patterns at three levels—by creator, by open-source status, and by model. Panel 14(a) shows competition among creators by reporting daily token shares by creator. Google and Anthropic together account for more than half of all usage until September 2025, but their relative positions shift frequently. Anthropic peaks at 50% in early 2025 following the release of Claude 3.5 and 3.7, before giving way to Google with the launch of Gemini 2.0 and 2.5, which push Google's share to just above 40% by mid-year. DeepSeek grows rapidly in mid-2025, reaching 20% at its peak, while OpenAI shows temporary gains around the introduction of GPT-4o. Notably, xAI has disrupted this dynamic in late August by gaining substantial market shares from Google and Anthropic following the release of its widely adopted model, Grok Code Fast 1. When we look at creators with smaller market shares, such as Meta-LLaMA, Mistral, and Qwen, they each maintain a 1–10% niche, and the residual "Others" category maintains a share of 2–15%, which fluctuated throughout the year. This figure highlights an industry in which usage is concentrated among a few large providers, yet leadership is constantly reshuffled with each new release.

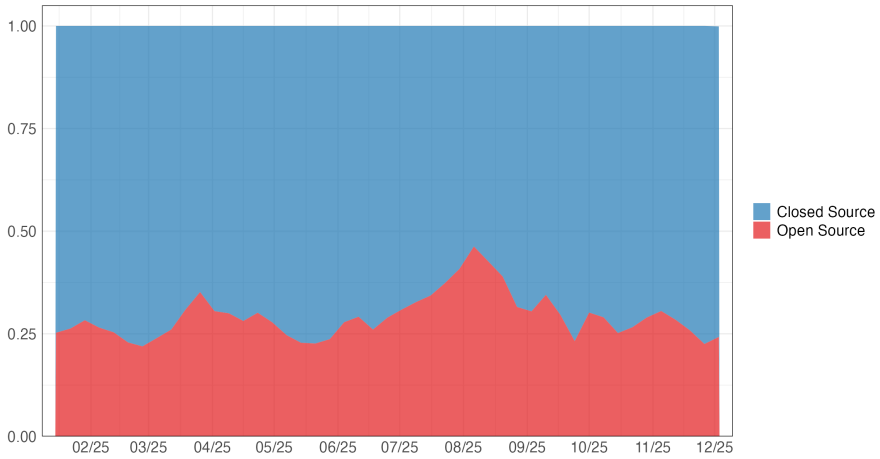
Figure 14(b) aggregates usage into open- and closed-source models. Closed-source models dominate, maintaining roughly 60%–75% of total tokens for most of the sample. Open-source models gained ground in the first half of 2025: their share rises from about 25% in January 2025 to 45% by August 2025. This increase is closely linked to the release of more capable open-source models, such as LLaMA 3, Mistral's dense models, and DeepSeek's efficient architectures, which have narrowed the performance gap while offering lower costs and greater flexibility. However, the market share for open-source models declined in the latter half of the year to 25% by December 2025, attributable to the emergence of popular closed-source xAI models such as Grok Code Fast 1. The figure suggests that although closed-source providers remain the leaders, there is competition for market share between open-source and closed-source models.

Figure 14(c) zooms in on the market shares of the top ten models as of December 2025. The figure reveals both concentration and substantial turnover. Although the top ten models

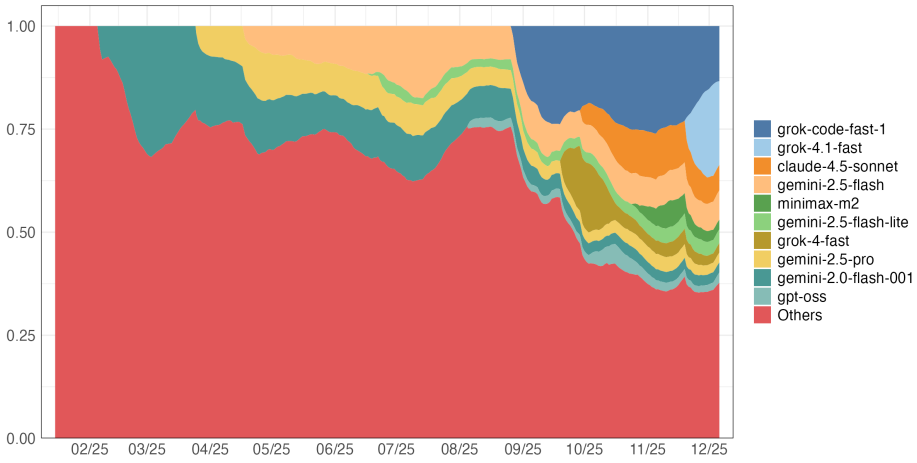
Figure 14: Aggregate Token Shares Over Time



(a) Daily Token Share by Creator



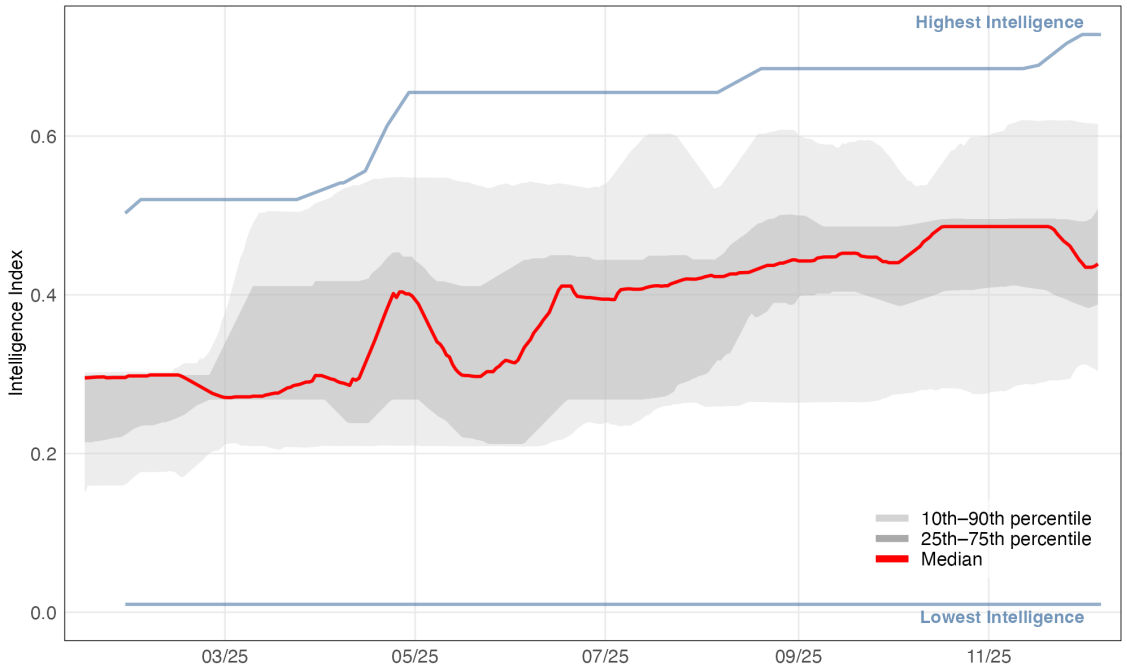
(b) Weekly Token Usage Share: Open vs. Closed-Source Models



(c) Daily Token Share of Top-10 Models

Notes: Subfigure 14(a) shows usage shares across model creators, Subfigure 14(b) aggregates usage between open- and closed-source models, and Subfigure 14(c) reports usage shares of the ten most widely used models. Rolling averages (14-day or weekly) are applied to smooth short-run variation. Market shares are based on the total number of token consumptions (completion + prompt).

Figure 15: Token-Weighted Distribution of Intelligence Over Time

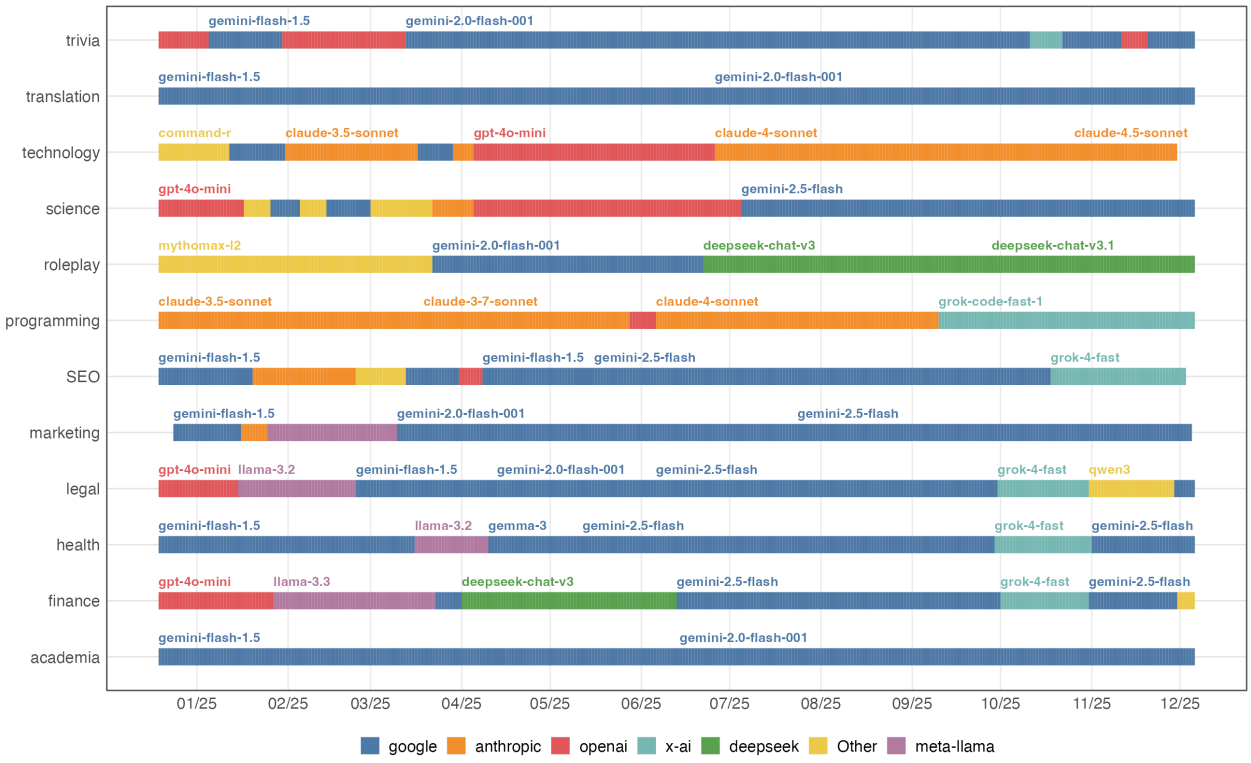


Notes: This figure shows the weighted distribution of model intelligences—measured using the Intelligence Index—weighted by total token usage over time. The sample includes all models available on OpenRouter. The red line represents the median intelligence of tokens used, the dark shaded area shows the 25th–75th percentile range, and the light shaded area shows the 10th–90th percentile range. The highest- and lowest-intelligence models are shown in blue, indicating the upper and lower bounds of intelligence observed in the sample.

collectively account for around 60% of total usage today, the same set accounted for only about 30% as recently as September 2025. Even more striking, these top ten models had zero market share in February 2025, as none had yet been released. Looking at individual models further underscores this rapid turnover. Models that dominated usage in early and mid-2025—such as Gemini 2.0 Flash and Gemini 2.5—now have negligible market shares. The panel overall highlights a dual feature of the market: strong responsiveness to innovation, with adoption shifting quickly after new releases, coupled with persistent concentration among a small set of leading models at any given point in time.

Next, we analyze the dynamics of the change in the intelligence of tokens over time in Figure 15. The figure plots the distribution of token-weighted intelligence over time—showing the median, interquartile range (25th–75th percentiles), and 10th–90th percentiles—alongside the lowest and highest intelligence levels available in the market, marked by red lines. We observe a steady upward shift in token-weighted intelligence, consistent with the introduction and diffusion of newer, more capable models. The median token's intelligence increases from 0.3 in January 2025 to 0.44 by year-end. Despite these increases, only a small fraction of tokens are associated with frontier-level intelligence. As of December 2025, the 90th-percentile token intelligence is only about 0.6, whereas the highest available intelligence

Figure 16: Category Market Leaders Over Time



Notes: Market leaders are based on 30-day rolling market share by weekly prompt tokens used for each category use case. The color indicates the creator of the top model in each category.

level exceeds 0.7. This persistent gap indicates that, despite rapid improvements at the technological frontier, most usage gravitates toward models whose lower prices more than offset their modestly lower intelligence—suggesting that the premium charged for frontier models may not be justified for many applications.

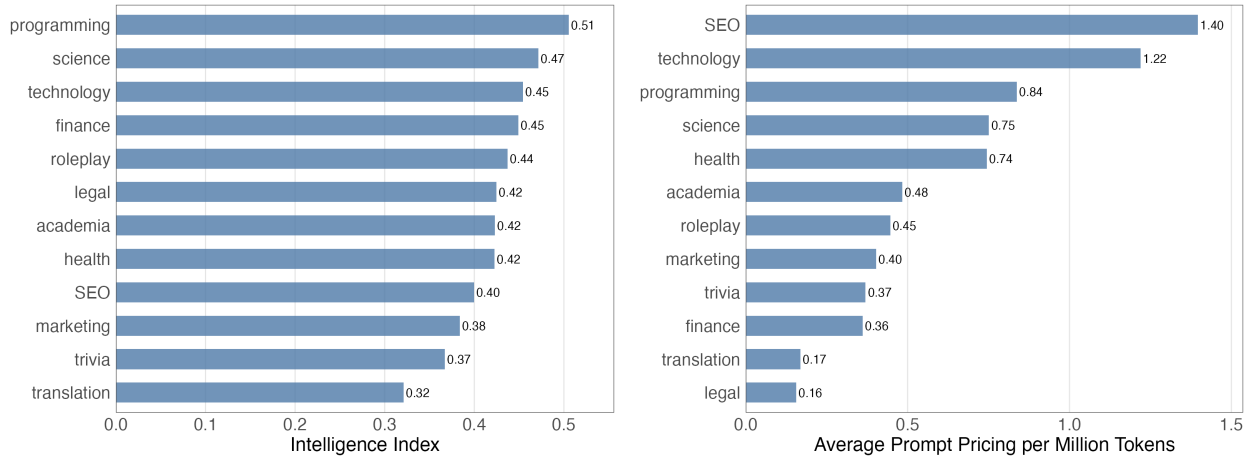
6.2 Demand for Intelligence Across Use Cases

One reason for the limited adoption of frontier models may be substantial heterogeneity in intelligence requirements across use case categories. To investigate this, we next analyze the dominant models within each major use case and the corresponding intelligence levels implied by their usage.

Figure 16 shows the leading model for each use case category. Colors denote the model's creator, while text labels indicate the specific model name. We highlight the top six creators—Google, Anthropic, OpenAI, DeepSeek, xAI, and Meta—and group all others into an "Other" label. The use cases include trivia, translation, technology, science, roleplay, programming, SEO, marketing, legal, health, finance, and academia.

Two clear patterns emerge. First, no single model dominates across all use cases. Sec-

Figure 17: Average Intelligence and Pricing by Use Case Category



(a) Average AI Intelligence Index **(b)** Average Prompt Pricing

Notes: Subfigure 17(a) shows the weighted average intelligence scores of the models used across categories over the 30-day period from November 7 to December 6, 2025. Subfigure 17(b) shows the average prompt price per million tokens over the same period across use case categories.

ond, model leadership is highly dynamic over time. Among current category leaders, Google models hold the top position in just over half of the categories. This underscores that different models may have comparative advantages in different domains—for example, Anthropic models were widely preferred by programmers until the introduction of xAI's Grok Code Fast 1. We also observe greater variation in category leaders earlier in the sample period, when models from LLaMA and OpenAI temporarily led in certain categories. This highlights both the market's competitive intensity and the horizontal differentiation among models.[17]

Another important result concerns the dynamics of the industry. This period has seen several major model launches, and across categories, the leading models are frequently replaced by new versions. In programming, for example, leadership transitioned rapidly from Claude 3.5 to Claude 3.7, then to Claude 4 and 4.5, immediately following their releases. Similarly, in categories where Google leads, we observe adoption of frontier models such as Gemini 2.5. However, the timing of adoption is not uniform across categories. Some, such as programming and legal, adopt newer models quickly, whereas others, such as trivia and marketing, have not yet transitioned to the latest version of their top model (Gemini 2.5).

A more direct way to observe this heterogeneity is to examine the average intelligence level across categories. In Figure 17(a), we report average intelligence scores over the 30-day period from November 7 to December 6, 2025. Programming emerges as the category with

[17] We also confirm heterogeneous model dominance across industries in the Microsoft data, as shown in Figure OA-5. Although the sample is primarily restricted to OpenAI models—because they are the only widely used models offered on Azure during our sample period—we still observe substantial variation across industries in the dominant models and their capabilities.

the highest average intelligence, scoring 0.51, followed by science, which scores 0.47. At the other end of the spectrum, trivia and translation show the lowest intelligence levels, with scores of 0.37 and 0.32, respectively.

The next question is how much users pay for this intelligence. Here, the heterogeneity is even more striking. As shown in Figure 17(b), the search engine optimization category pays an average of $1.40 per million prompt tokens, while translation pays just $0.17 and legal a mere $0.16.

Taken together, these patterns illustrate a sharp divergence in both the demand for intelligence and the willingness to pay across use cases. High-stakes domains such as search engine optimization, programming, and technology place substantial value on incremental improvements in capability and are willing to pay a premium for frontier models. By contrast, in domains where tasks are simpler or less sensitive to model intelligence—such as trivia or translation—users gravitate toward lower-priced options, with little incentive to adopt the most advanced models.

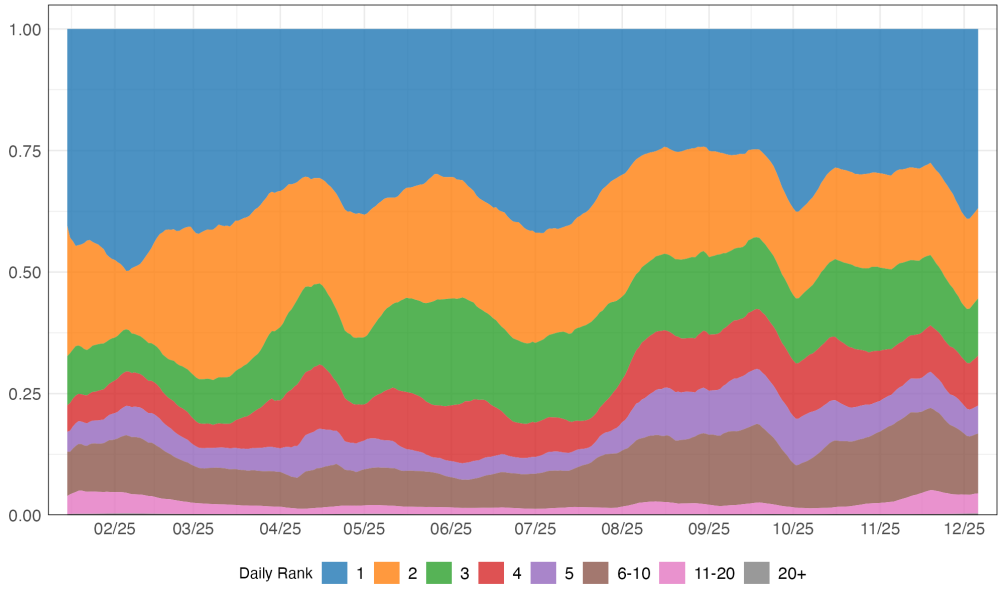
In other words, vertical differentiation in quality interacts with horizontal differentiation in use cases. The market is not one in which a single "best" model dominates across all tasks. Instead, model choice is mediated by the extent to which each use case values incremental intelligence relative to cost. This heterogeneity in demand is a key reason why multiple creators remain competitive, and why leadership is both dynamic and fragmented across categories.

6.3 Concentration

Next, we directly analyze market concentration on the OpenRouter platform by creator. Figure 18 reports market shares by rank, focusing on the top five creators and grouping others into ranks 6–10, 11–20, and 20+. A few observations stand out. First, overall concentration appears relatively stable over the sample period, with no evidence of a single creator dominating the market. Day-to-day fluctuations occur, but the largest creator holds around 25%–50% of usage, while the second-ranked creator ranges between 15%–30%. Beyond the top two, other leading creators also maintain significant shares: creators ranked 6–10 collectively account for around 5%-15%. Even providers ranked outside the top 10 capture several percentage points, suggesting meaningful competitive pressure from smaller players.

In Figure 19, we analyze concentration more formally using the Herfindahl-Hirschman Index (HHI). The HHI is a standard measure of market concentration used in antitrust analysis, calculated as the sum of squared market shares. Values range from 10,000 in a monopoly to 0 in a perfectly competitive market, with intermediate values indicating the degree of concentration (for example, an HHI of 2,500 is equivalent to four equally sized

Figure 18: Token Usage Share by Creator Rank over Time



Notes: This figure reports creator market shares by rank from January 2025 through December 2025, highlighting the top five creators and grouping the remainder into ranks 6–10, 11–20, and 20+. A rank of 2, for instance, denotes the market share of the second-largest creator on that day, which varies over time. The figure presents a 14-day rolling average, and market shares are computed from total token consumption (prompt and completion tokens combined).

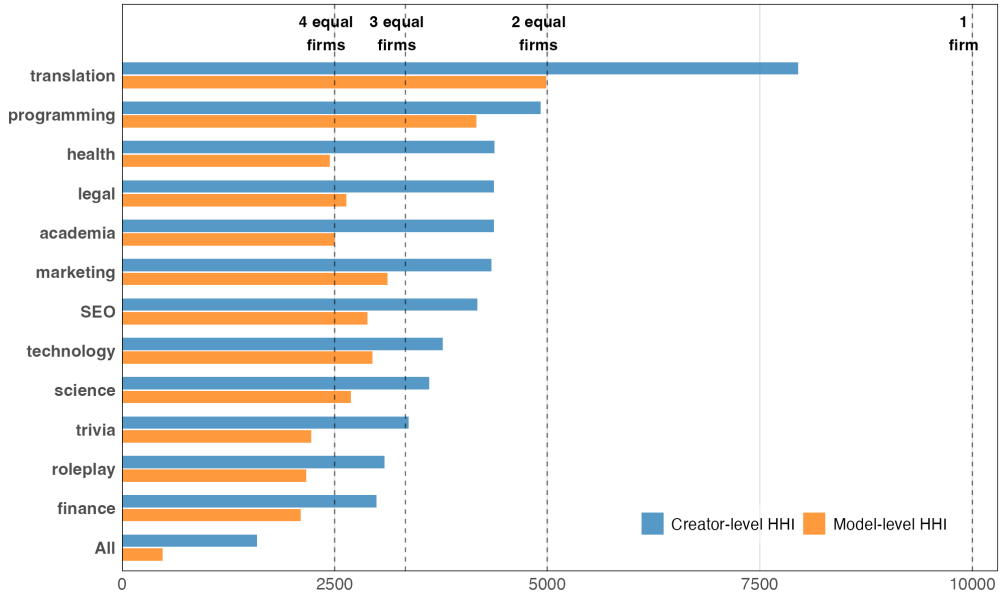
firms). The figure shows HHIs at both the model level (orange) and creator level (blue) across categories, along with benchmark lines for interpreting concentration levels.[18]

The results show that concentration is high across most categories. Translation is the most concentrated, with HHIs approaching 8,000—equivalent to a duopoly market with asymmetric firms. Many other categories, including programming, health, and marketing, record HHIs between 3,000 and 5,000, equivalent to a concentration level when only two or three firms of equal size compete. Comparing model- and creator-level HHIs reveals that they do not differ significantly from each other, indicating that when multiple models are used in a given category, they often originate from the same creator. Importantly, concentration levels do not map directly onto willingness to pay for quality. Earlier, we found that programming and technology exhibit the highest willingness to pay, whereas translation exhibits the lowest. Yet translation is the most concentrated category, showing that competitive structure is not tightly coupled to demand-side valuation.

Finally, the overall market concentration ("All" category) is significantly lower than the concentration within individual categories. This indicates that while certain models dominate in specific domains, multiple creators and models are active across the broader set of use

[18] Our decision to analyze concentration at the category level should not be interpreted as a claim that each category constitutes a relevant antitrust market. Rather, our goal is simply to examine usage patterns within categories and compare concentration across them.

Figure 19: Average Market Concentration by Category (HHI Index)



Notes: The Herfindahl-Hirschman Index (HHI) is a standard measure of market concentration used in antitrust analysis, calculated as the sum of squared market shares. Values range from 10,000 in a monopoly (one firm) to 2,500 in a market with four equally sized firms. The figure shows HHIs at both the model level (orange) and creator level (blue) across categories, along with benchmark lines for interpreting concentration levels. Market shares are based on total token usage. The sample period is from January 2025 to December 2025.

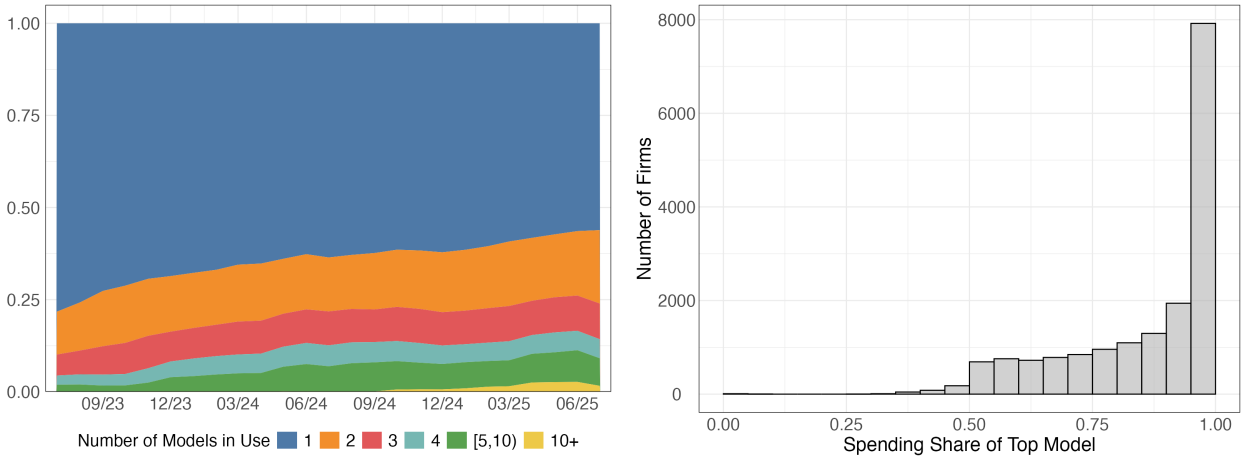
cases. Thus, competition is more fragmented at the aggregate level, even though categories individually tend to be more concentrated - implying that dominance is category specific rather than market wide.

6.4 Multihoming

The final set of facts we document concerns the extent to which firms multi-home, i.e., use multiple models within a given period. To study this, we analyze granular model-level usage data from Microsoft Azure, enabling us to observe firms' adoption patterns across multiple LLMs. In Figure 20(a), we present the share of firms using multiple models over time, measured at the monthly level, distinguishing between those using 1, 2, 3, 4, 5–10, and 10+ models. The plots indicate that most firms use a single model for all API queries. However, this share has declined significantly since mid-2023 (from over 75% to slightly more than 50%), as the share of firms using 2-5 models has risen steadily over the same period. A much smaller share but growing share of firms in the right tail employ more than five distinct models, suggesting experimentation or use case segmentation.

One limitation of these simple multi-homing measures is that they do not capture the intensive margin of model usage. A firm may rely heavily on a single model while experimenting with others, or shift workloads between models across periods. To account for this, Figure 20(b) reports the distribution of usage shares across a firm's top models, conditional

Figure 20: Patterns of Multihoming Across Firms



(a) Share of Firms by Number of Models Used Over Time

(b) Distribution of Usage Shares Across Multihoming Firms' Top Model

Notes: Subfigure 20(a) shows the share of firms using 1, 2, 3, 4, 5–10, or 10+ models in a month over time. Subfigure 20(b) reports the distribution of usage shares across multihoming firms' top models as of June 2025, conditional on firms using at least two models in June 2025.

on multi-homing (using at least two models) as of June 2025. We find that, even among these multi-homing firms, most spend nearly all of their consumption on a single model, though a substantial share also use a second model for at least 15% of usage. Our interpretation of these patterns is that, for most firms, multi-homing reflects experimentation rather than intensive dual use of models tailored to specific tasks.

7 Demand: Price Elasticities and Differentiation Among Providers

In this section, we examine how demand for AI models responds to changes in price, performance, and other characteristics. Understanding these demand relationships is crucial for two reasons. First, it helps predict how the AI industry will evolve as models become more efficient. Second, it speaks directly to a prominent debate about the so-called Jevons Paradox in AI—the possibility that efficiency improvements could increase total resource consumption.

7.1 Theoretical Framework

Jevons' Paradox occurs when technological improvements that reduce the cost of using a resource lead to such significant increases in demand that total resource consumption rises rather than falls. In the context of AI, this would mean that more efficient models, requiring less compute per task, could increase total compute and energy usage.

The paradox emerges under specific economic conditions. Consider an aggregate demand

function $D(p)$ and a competitive supply with constant marginal cost c, such that $p = c$. Total expenditure equals $p \cdot D(p) = c \cdot D(c)$. Jevons Paradox occurs when a reduction in c increases total expenditure, which requires that the price elasticity of demand exceeds unity in absolute value: $|\epsilon_D| = |d \log D / d \log p| > 1$.

While this aggregate condition provides the benchmark, demand estimates are typically made at a lower level of aggregation. We estimate demand at two levels: (1) across models, capturing how users substitute between different AI models, and (2) across providers of the same model. Model-provider level elasticities are larger in magnitude than the corresponding aggregate elasticity because they measure substitution across models and providers rather than solely between the outside option and the set of models and providers.

7.2 Empirical Strategy

The primary challenge in estimating demand elasticities is the presence of price endogeneity. High-quality models command both higher prices and higher quantities, potentially biasing naive price coefficients upward. We address this challenge through a within-model identification strategy that uses variation in prices across providers offering the same model.

Specifically, we estimate:

$$\log(Q_{imt}) = \beta_1 \log(\text{Price}_{imt}) + \beta_2 \log(\text{Throughput}_{imt}) + \beta_3 \log(\text{Latency}_{imt}) \tag{1}$$

$$+ \beta_4 \log(\text{Context}_{imt}) + \gamma_t m + \theta_i m + \varepsilon_{imt} \tag{2}$$

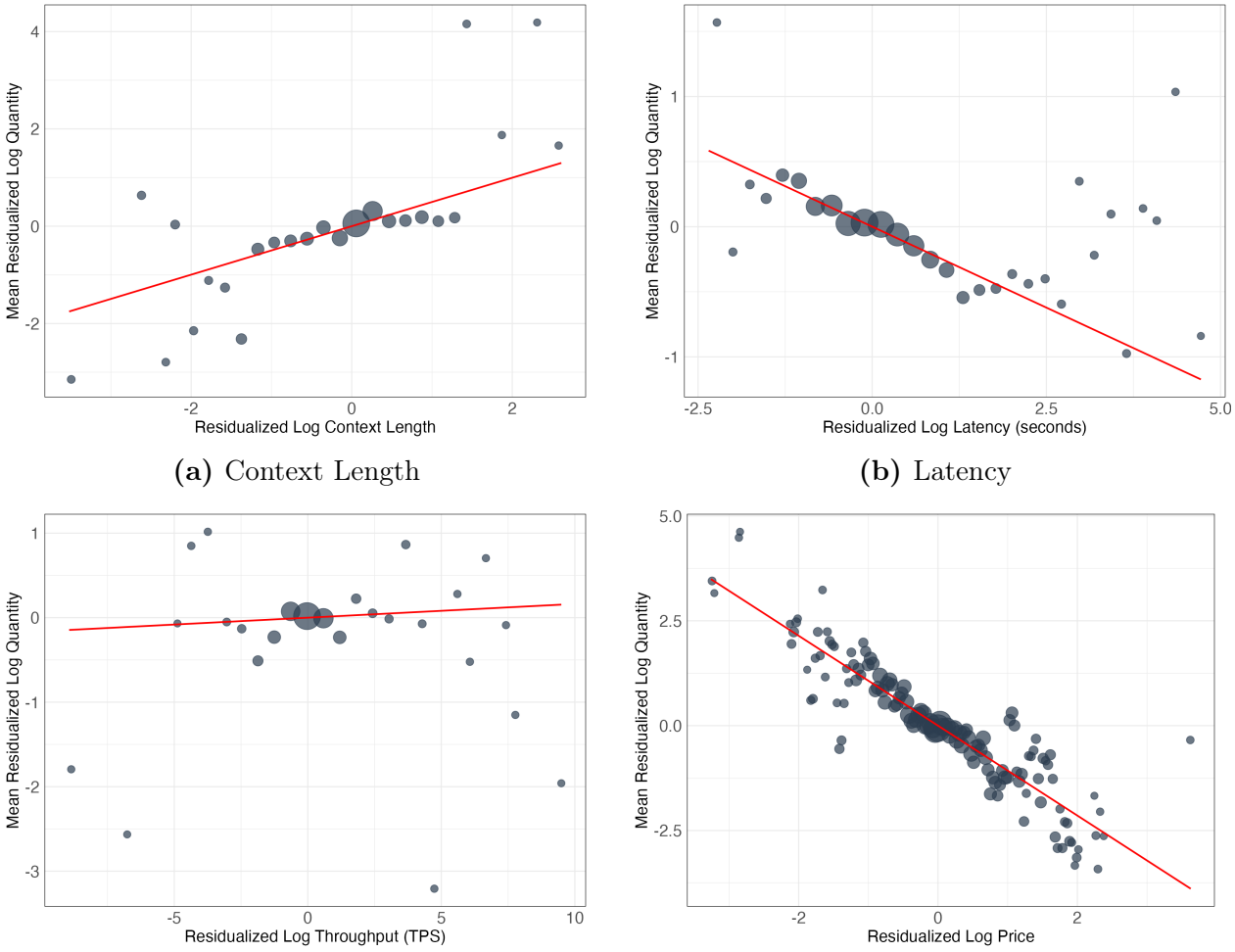
where Q_{imt} denotes daily total tokens for provider i offering model m on date t. The key covariates capture:

- Price_{imt}: prompt price per token
- Throughput_{imt}: inference speed (tokens/second)
- Latency_{imt}: time to first token (seconds)
- Context_{imt}: maximum context window length

The fixed effects γ_{tm} and θ_{im} control for time by model trends and provider-model quality, respectively. We also experiment with alternative fixed-effect configurations.

Our identification of β_1 relies on within-model price variation arising from two sources: (1) entry and exit of providers offering the same model, and (2) price changes by existing providers. Since model quality is held constant, this variation plausibly isolates the causal effect of price on quantity demanded. Our data for this estimation comprises just open-source models, since these exhibit within-model price variation.

Figure 21: Residualized Quantity vs. Provider Attributes for Open-Source Models



(a) Context Length

(b) Latency

(c) Throughput

(d) Price

Notes: Bin scatter plots of residualized log total token usage against provider attributes for open-source models. Subfigure 21(a) shows context length; Subfigure 21(b) shows latency (seconds); Subfigure 21(c) shows throughput (transactions per second); Subfigure 21(d) shows price. The x-axes report residualized log values of provider attributes, and the y-axes report mean residualized log total token usage. The unit of observation is a provider–model-day in the analysis; fitted regression lines are shown in red. In each plot, the variables are residualized with respect to the other three attributes in other plots.

7.3 Empirical Results

Figure 21 presents bin scatter plots of residualized log quantity against four key attributes: price, latency, context length, and throughput. These bin scatters are based on residualized variables, meaning that the effects of other attributes, as well as model and date fixed effects, have been partialled out before the relationship is plotted.

The plots show patterns consistent with higher quality being associated with greater usage. Specifically, longer context lengths and higher throughput are positively associated with usage, while higher latency and higher prices are negatively associated with usage. In

Table 2: Price Elasticity Regressions

	Log(Daily Tokens)		
	(1)	(2)	(3)
Log(Price)	-0.55***	-1.08***	-1.11***
	(0.09)	(0.19)	(0.22)
Log(Throughput)	-0.40**	-0.01	-0.07
	(0.16)	(0.08)	(0.05)
Log(Latency)	-0.11	-0.31***	-0.13***
	(0.24)	(0.08)	(0.04)
Log(Context Length)	0.82***	0.27***	0.22
	(0.13)	(0.10)	(0.24)
Observations	32,539	32,539	32,539
R^2	0.34	0.76	0.97
Within R^2	0.31	0.16	0.06
Date fixed effects	✓	✓	
Model fixed effects		✓	
Date × Model fixed effects			✓
Model × Provider fixed effects			✓

Notes: This table reports regression results estimating price elasticity of demand at the provider-model-day level. The dependent variable is the log quantity (tokens). The main independent variable is log price (per million tokens). Controls include provider performance characteristics (context window ratio, throughput ratio, latency ratio), model characteristics (Intelligence Index, open source indicator), and temporal effects. All specifications include model- and provider fixed effects. Standard errors clustered at the provider level are reported in parentheses.

terms of magnitudes, the strongest relationship is observed with price.[19]

Table 2 displays the demand model estimates. Column (1) includes the above variables and just a date fixed effect. This coefficient on price in this specification is -0.55, suggesting a very small elasticity; however, without accounting for differences in model quality this estimate is biased. In Column (2), we add model fixed effects, allowing us to control for unobserved quality differences across models. Under this specification, we find a more sizable elasticity, -1.08, which is at least potentially consistent with Jevons' paradox. However, it may be that cheaper providers of a given model are also better in other respects, not captured by our covariates. In column (3), we add provider-model and date-model fixed effects. Identification now derives solely from variation within day and model. We find an elasticity of -1.11.

We also find effects consistent with the expected signs for latency and context length. When these increase, the number of daily tokens decreases. However, in our preferred spec-

[19] An important caveat is that OpenRouter allows users to either select a specific provider or delegate the choice to its routing algorithm. OpenRouter's algorithm selects providers based on a combination of price and other attributes. Thus, some of the observed price sensitivity may reflect routing decisions made by OpenRouter rather than direct user choices.

ification (3), higher throughput is associated with lower demand. This is consistent with providers having finite capacity. When more tokens are requested, throughput decreases. Note that this raises endogeneity concerns, since demand affects throughput. In Table OA-5, we show a version of the demand estimation without provider performance metrics and find very similar results.

Although the elasticities we find are consistent with a provider-level Jevons paradox effect, they are very close to one in magnitude. Since aggregate demand elasticities are substantially smaller, we interpret our findings as going against the Jevons paradox in the short-run. This, however, does not preclude from Jevons effects operating on longer-run horizons. Firms may take time to decide whether and to what extent to use LLMs. As a result, short-run price fluctuations may not lead to significant changes in quantity, even as year-over-year price declines result in substantially more LLM usage. This is consistent with the huge growth in tokens served across OpenRouter and the industry as a whole.

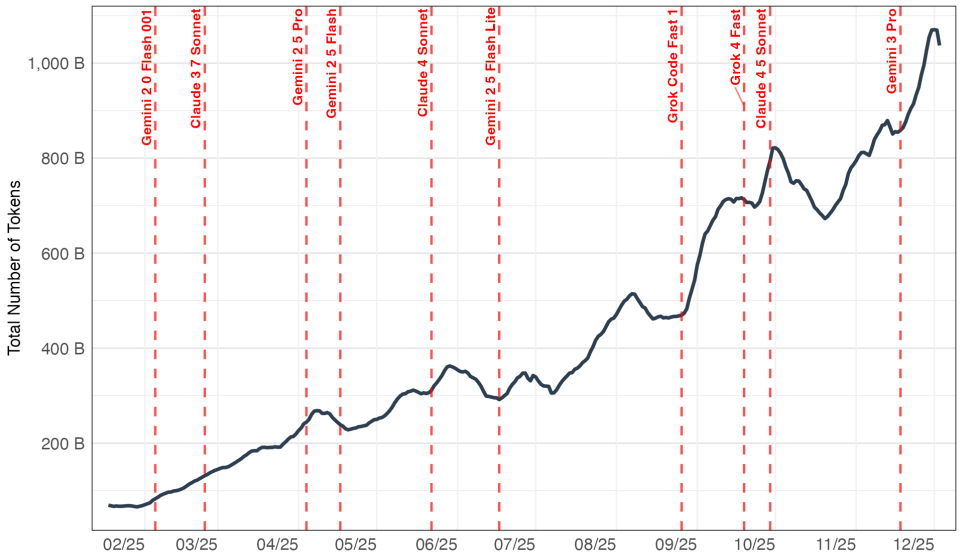
8 Substitution and Market Expansion: Evidence from Model Entry

In this section, we examine how new model releases affect the demand for existing models. Understanding substitution patterns is important for characterizing competition in the intelligence market: do new models primarily cannibalize their predecessors, steal share from rivals, or expand the overall market? We conduct a series of case studies using data from OpenRouter and Microsoft Azure, focusing on major model releases and tracking usage of competing models in the days surrounding each launch. We find heterogeneous substitution patterns across providers. Some model families exhibit strong within-brand cannibalization, while others appear to expand the market with minimal displacement of competitors. We also show that users of a given Coding app demand different models at any given point in time. These patterns corroborate our earlier evidence on differential model demand across use cases, and suggest that models are differentiated in ways not fully captured by public benchmarks.

The ideal experiment to measure substitution would be to randomly assign model access to some users and not others. Unfortunately, we do not have access to such an experiment. Instead, we conduct an interrupted time-series analysis. This requires many assumptions, two of which are particularly important. First, there should be no concurrent events that shock the market when a model becomes available. This assumption may be violated if, for example, a new developer begins using OpenRouter concurrently with the launch of a new model. The second assumption is that the model entry does not mechanically induce users to switch from a deprecated model on OpenRouter to the new model.

If new model releases increase token demand substantially, this should be evident in the

Figure 22: Total Consumed Tokens and Model Entry



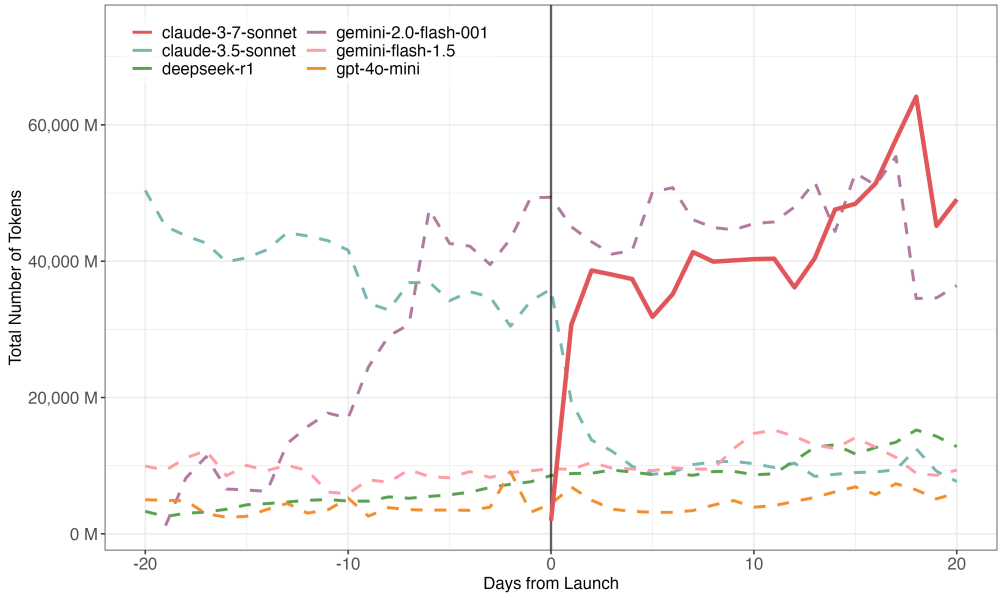
Notes: This figure shows the evolution of the 7-day rolling average of tokens over time with vertical lines indicating major model releases on OpenRouter. The x-axis is calendar time, and the y-axis shows the total number of tokens. Model entry dates are marked for official, non-experimental releases of major models.

time series; Figure 22 shows total tokens over time, with vertical lines denoting major model entries on OpenRouter. We define entry as the time when the official, non-experimental, non-secret version of the model is released.[20] We see that new model releases do not seem to cause abnormally great changes in token demand relative to the overall trend (with the exception of the release of Grok Code Fast 1, which we discuss further below). One way to interpret this is that, relative to the overall secular trend in token demand, the effects of individual models during this period are small.

Next, we examine substitution across models upon the release of a new model, both using OpenRouter and Microsoft data. For OpenRouter case studies, we consider models that achieved substantial success on OpenRouter at the time of entry. These include: Claude 3.7 Sonnet, Claude 4 Sonnet, Gemini 2.0 Flash, Gemini 2.5 Flash, Gemini 2.5 Pro, and Grok Code Fast 1. For each entering model, we select the five most popular models at the time of entry as comparison models. We then plot the 20 days around the entry. In the Microsoft case studies, we analyze the introduction of two models—GPT-4o and DeepSeek R1—and measure substitution as the change in the share of firms that use existing models. By comparing trends across these comparison models on both platforms, we can assess whether any anomalous shifts in demand are plausibly attributable to the focal model's release.

[20]For example, we exclude Gemini 2.0 Flash Experimental (free), which was available in December 2024 but capacity-constrained. We also exclude "Optimus Alpha," a version of OpenAI's GPT-4.1 that was temporarily offered for free on OpenRouter prior to the official release of 4.1.

Figure 23: Usage for Select Models Following Claude 3.7 Sonnet Release



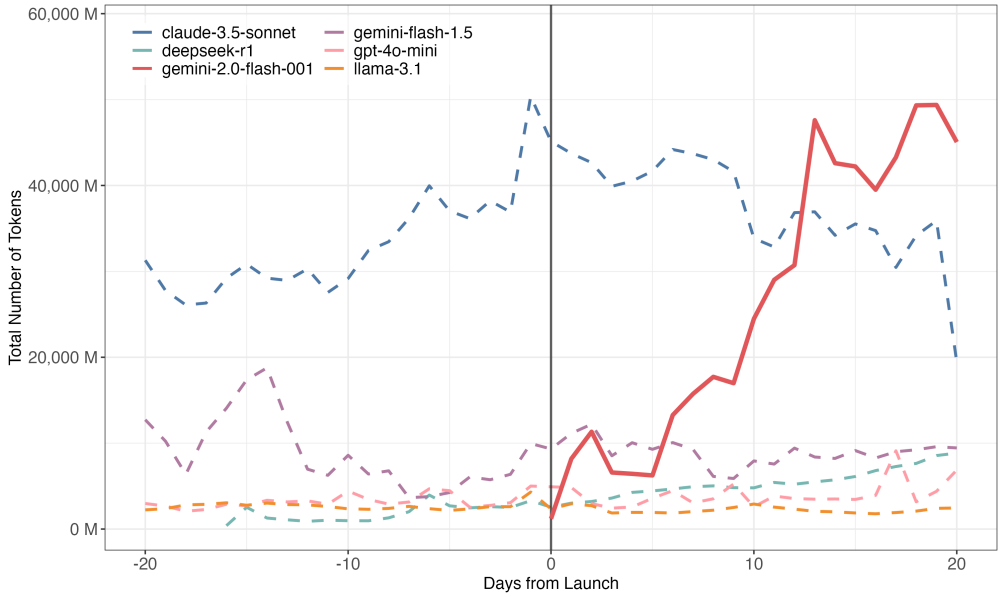
Notes: Daily token usage for selected models around the Claude 3.7 Sonnet launch. The x-axis is labeled in days from launch (vertical black line at day 0), spanning approximately -20 to +20 days. Series shown: `claude-3.7-sonnet`, `claude-3.5-sonnet`, `deepseek-chat-v3`, `gemini-2.0-flash-001`, `gpt-4o`, `llama-3.1`. y-axis units are millions of tokens.

8.1 Case Studies from Open Router

Claude 3.7 Release: Claude 3.7 Sonnet was released by Anthropic on Feb 24, 2025. At the time of its release, Claude 3.7 Sonnet was a state-of-the-art model, marketed for its strengths in coding and front-end development. Figure 23 plots the tokens used by the model relative to the launch date. It reveals a striking pattern of within-creator substitution. Upon the release of Claude 3.7 Sonnet, we observe an immediate and substantial decline in the usage of Claude 3.5 Sonnet, suggesting strong substitutability between successive generations. The decline in Claude 3.5 usage is nearly one-for-one with the uptake of Claude 3.7, indicating that users view these models as close substitutes rather than complements. Interestingly, we observe minimal impact on other major models, such as GPT-4o and Gemini 2.0 Flash, suggesting that Claude's competitive positioning is primarily within its own model family rather than across providers.

Claude 4 Sonnet Release: Claude 4 Sonnet was another major release by Anthropic, improving upon the benchmarks of Claude 3.7 Sonnet. Similar to the Claude 3.7 release, Appendix Figure OA-16 shows within-family substitution. The entry of Claude 4 Sonnet results in an immediate cannibalization of Claude 3.7 Sonnet usage, with the magnitude of the decline in the predecessor model closely matching the uptake of the new model. Interestingly, demand for Claude 3.7 remains substantial even after the introduction of Claude 4 Sonnet. This concurrent demand for Claude 3.7 and 4 provides evidence that models are differentiated

Figure 24: Usage for Select Models Following Gemini 2.0 Flash Release



Notes: Daily token usage for selected models around the Gemini 2.0 Flash launch (February 5, 2025). The x-axis is days from launch with a vertical black line at day 0 (window $\approx$ -20 to +20 days); the y-axis reports tokens. Series include `gemini-2.0-flash-001` (focal) and contemporaneous leading models.

in ways that are not captured by public benchmarks.

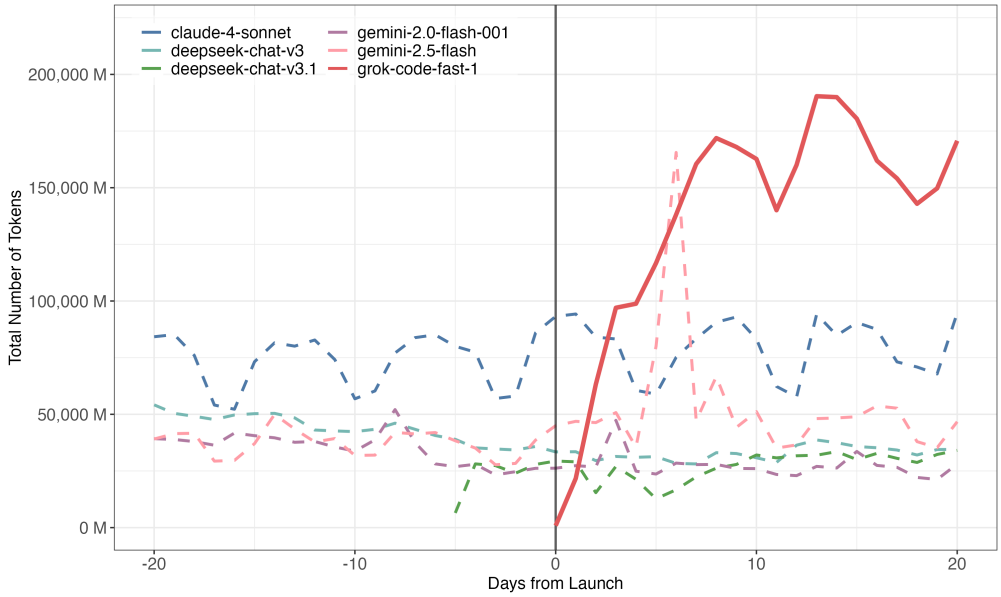
Gemini 2.0 Flash Release: Google's Gemini 2.0 Flash was released on February 5, 2025. The model lacked frontier capabilities at the time of its release, but it was fast and cost-effective. In contrast to the Claude releases, Figure 24 reveals a different substitution pattern. Rather than the clean within-family substitution observed with Claude models, Gemini 2.0 Flash's entry appears to expand the market, with overall token demand growth tracking Gemini 2.0 Flash usage.

Gemini 2.5 Flash Release: The Gemini 2.5 Flash model made similar tradeoffs to Gemini 2.0 Flash; it was fast and cheap, but not a frontier model. Figure OA-17 in the Appendix continues the pattern observed with Gemini 2.0 Flash, showing few if any visible substitution effects with other popular models at the time of the release.

Gemini 2.5 Pro Release: Gemini 2.5 Pro was Google's flagship model in the 2.5 series, designed for complex reasoning and professional applications. This model was considered frontier at the time of release. Appendix Figure OA-18 shows that Gemini 2.5 pro gains market share over time as demand for DeepSeek Chat V3 and Claude 3.7 flattens. Note that an experimental, rate-limited version of Gemini 2.5 Pro was released approximately 10 days prior to Gemini 2.5 Pro's full launch. There is no tell-tale discontinuity at the time of Gemini 2.5 Pro's full or experimental release to suggest substitution.

Grok Code Fast 1: xAI's Grok Code Fast 1 was released in late August 2025 and was billed as xAI's fast and cheap competitor to existing coding models. As shown in Figure

Figure 25: Usage for Select Models Following Grok Code Fast 1 Release



Notes: Daily token usage for selected models around the Grok Code Fast 1 launch. The x-axis is days from launch with a vertical black line at day 0 (window $\approx$ -20 to +20 days); the y-axis reports tokens. Series include `grok-code-fast-1` (focal) and contemporaneous leading models.

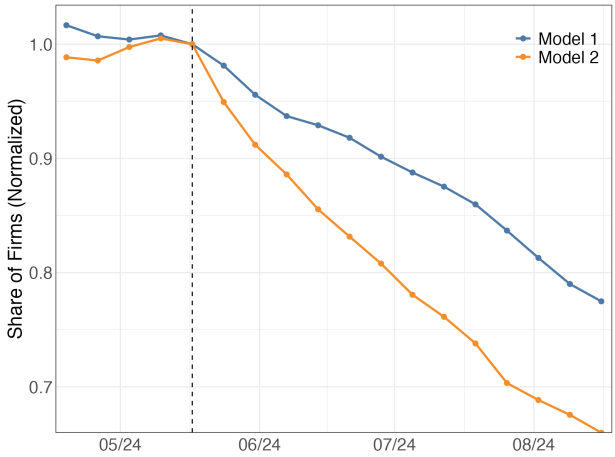
OA-17, it achieved a huge market share within a week of launch. This was done through aggressive partnerships with coding apps and subsidizing pricing. For example, this model was free on Roo Code, a popular open-source coding app. Perhaps surprisingly, given the success of Grok, the other top models during this period show no change in their demand trends. For example, Claude 4 Sonnet, which was perhaps the premier coding LLM at the time, continues on a steady trend around the time of the launch. We conclude that Grok Code Fast 1 increased the market for LLMs upon entry.

8.2 Case Studies from Microsoft Azure

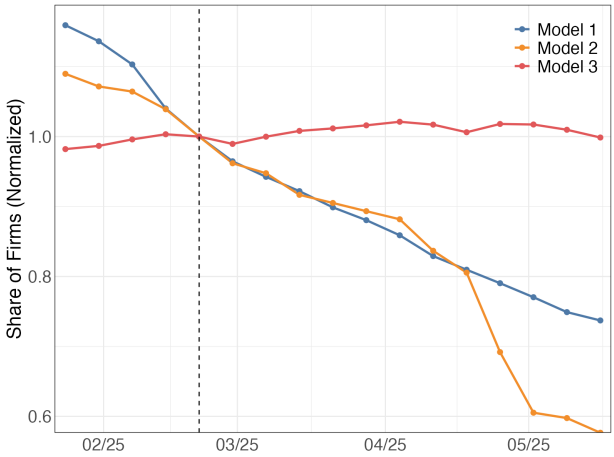
GPT-4o Release: OpenAI's GPT-4o model was released in May 2024, and represented a significant improvement over GPT-3.5 and GPT-4 when launched. In Figure 26(a) we plot the share of firms using each of the top two most popular models on Azure's API at the time of GPT-4o's launch. We normalize each share to its value during the GPT-4o launch week. Here we see an immediate and steep decline in the share of customers using other popular models after GPT-4o's introduction, while the usage trends of these two popular models were relatively flat leading up to the introduction of GPT-4o.

Deepseek R1 Release: Deepseek's R1 model was released in late May, 2025. At the time, this model received substantial attention, in part because it was an open-weight model that performed comparably to some closed-source frontier models. In Figure 26(b), we again show the normalized share of firms using top models at the time. Interestingly, the results here are

Figure 26: Firm-Level Usage Following the Release of GPT-4o and Deepseek R1



(a) Change in Other Model Usage after GPT-4o Introduction



(b) Change in Other Model Usage after Deepseek R-1 Introduction

Notes: Subfigure 26(a) shows the normalized share of firms using the top two most popular models on Azure's API at the time of GPT-4o's launch (May 2024). Subfigure 26(b) shows the normalized share of firms using the top three most popular models at the time of Deepseek R1's launch (late May 2025). Each share is normalized to its value during the launch week.

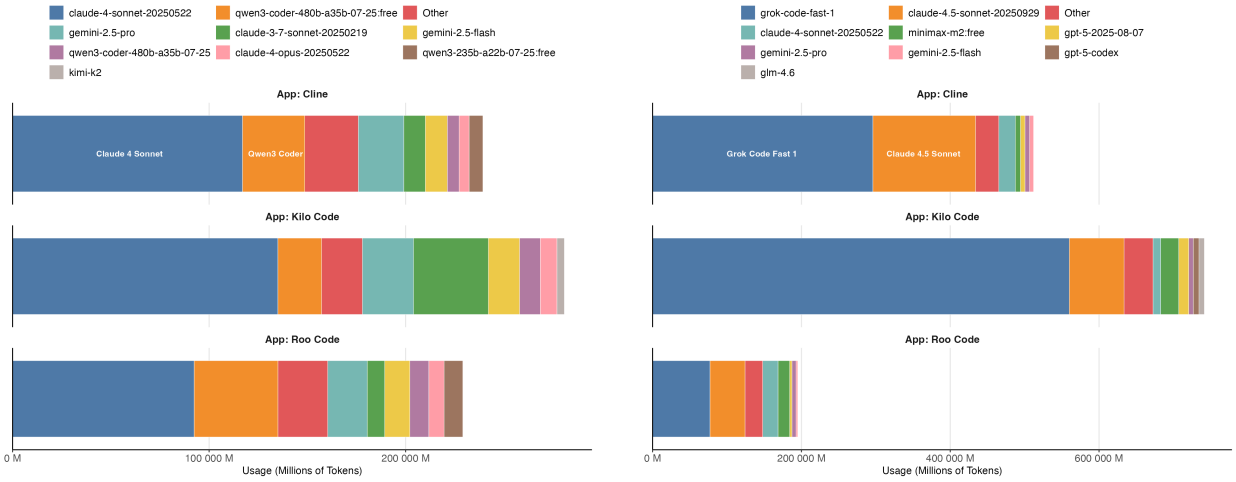
in contrast to the impact of GPT-4o. For all three of the most popular models, we do not observe any apparent change in usage trends for the three most popular models hosted by Microsoft before and after R1 was released. However, we cannot rule out responses among smaller models or among models not hosted by Microsoft.

Synthesis: Heterogeneous Preferences and Substitution Patterns Our case studies reveal differences in how models from different providers compete in the intelligence market. There are two distinct substitution and market expansion patterns:

Brand-Centric Competition (Claude): Anthropic's Claude models exhibit within-family substitution, where new releases cannibalize predecessors with minimal cross-provider effects. This pattern suggests two potential explanations. First, Anthropic has established a brand identity in which users exhibit strong loyalty and view successive model generations as close substitutes. Alternatively, Anthropic's models may be better in ways not captured by publicly available benchmarks that users care about. The fast substitution patterns also indicate low switching costs within the Claude ecosystem.

Performance-Centric Competition and Market Expansion (Gemini Flash and Grok Code Fast): Google's Gemini models demonstrate broader, more diffuse substitution patterns. Of particular note is that the Gemini 2.0 Flash model gained substantial market share without causing noticeable declines in the usage of other models. This suggests that the model was selected at a price-to-performance point previously unoccupied by other

Figure 27: Model Usage Distribution for Popular Coding Apps



(a) July 30, 2025 **(b)** October 30, 2025

Notes: Subfigure 27(a) shows the distribution of tokens used on July 30, 2025 for three popular coding applications on OpenRouter. Similarly, Subfigure 27(b) shows the distribution of tokens used on October 30, 2025

models. Other Gemini models achieved moderate success, without either greatly expanding the market or causing large-scale substitution from any specific model. Similar to Google's Flash models, Grok Code Fast 1 achieved substantial market share without causing noticeable declines in the demand for other models.

8.3 App-Specific Usage Patterns and Alternative Explanations

The preceding analysis considered aggregate token demand and substitution patterns across models. However, substitution can arise from several sources: firms actively choosing models for internal AI applications, the aggregated choices of end users of consumer-facing apps, or decisions by OpenRouter or app developers to steer their users toward particular models. In this section, we discuss these explanations in greater detail by using examples from individual coding applications.

Figure 27 displays the distribution of token usage across three popular open-source VS-Code plugins—Cline, Kilo Code, and Roo Code—at two points in time. Users of these apps are individual developers who select models through a simple interface. Between July and October 2025, the most popular model across all three apps shifted from Claude 4 Sonnet to Grok Code Fast 1, while most remaining Sonnet usage migrated to Claude 4.5 Sonnet. The figures also reveal a substantial long tail of model usage, consistent with heterogeneous user preferences.

However, an alternative explanation for these shifts is a default or salience effect. Cod-

ing apps can set recommendations and defaults, and may have financial incentives to promote certain models. The routing method—whether through OpenRouter or directly via a provider—may also depend on app-level choices. Notably, all three apps feature Grok Code Fast 1 prominently as a free option. Thus, at least some of the switching we observe may reflect default settings or promotional placement rather than informed user choice.

9 Conclusion

This paper uses large-scale marketplace data from two LLM API platforms, OpenRouter and Microsoft Azure Foundry, to measure how the market for LLMs has evolved - characterizing supply, pricing, demand, and usage dynamics since mid-2023. The supply side is characterized by rapid entry, particularly from open-source models and inference providers, and by sharp declines in intelligence-adjusted token prices. We find that large price dispersion persists even among models with similar benchmark performance; open-source models are 90% cheaper than comparable closed-source models, yet account for less than 30% of the market on average, indicating meaningful non-price differentiation.

Demand is similarly dynamic and heterogeneous. Market-share leadership shifts frequently across both models and creators, and no single model dominates across use cases. The willingness to pay for incremental intelligence varies widely across applications, and most API usage remains below the frontier. We find evidence of heterogeneity in substitution patterns following new model entries, with some entrants cannibalizing predecessors within a model family, while others gain share with little displacement of contemporaneous competitors - patterns consistent with meaningful horizontal differentiation. Finally, using within-model provider price variation, we estimate price elasticities that are inconsistent with model- or market-level Jevons paradox effects in the short run.

These findings reveal a market in rapid transition, with important implications for the future of AI adoption and competition. The proliferation of inference providers—particularly for open-source models—has created competitive pressures that benefit users by lowering prices and increasing choice. Of particular interest for future research are longer-run adjustments by firms and other users to these lower prices and greater capabilities.

References

Acemoglu, D. (2024). The Simple Macroeconomics of AI. *NBER Working Paper* (32487).

Acemoglu, D. and P. Restrepo (2018). Artificial Intelligence, Automation, and Work. In *The Economics of Artificial Intelligence: An Agenda*, pp. 197–236. University of Chicago Press.

Acemoglu, D. and P. Restrepo (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives 33*(2), 3–30.

Andreessen Horowitz (2025). How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025. Technical report, Andreessen Horowitz.

Aubakirova, M. and A. Midha (2025). State of AI: An Empirical 100 Trillion Token Study with OpenRouter. Technical report, Andreessen Horowitz.

Autor, D. and N. Thompson (2025). Expertise. *Journal of the European Economic Association 23*(4), 1203–1271.

Bick, A., A. Blandin, and D. J. Deming (2024). The Rapid Adoption of Generative AI. *NBER Working Paper* (32966).

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at Work. *The Quarterly Journal of Economics 140*(2), 889–942.

Brynjolfsson, E., T. Mitchell, and D. Rock (2018). What Can Machines Learn and What Does it Mean for Occupations and the Economy? *AEA Papers and Proceedings 108*, 43–47.

Chatterji, A., T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman (2025). How People Use ChatGPT. *NBER Working Paper* (34255).

Cui, Z. K., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2025). The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Available at SSRN* (4945566).

Dell'Acqua, F., R. Agarwal, M. Iansiti, and S. Zheng (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of The Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24-013).

Deloitte (2024). State of Generative AI in the Enterprise Quarter Three Report. Technical report, Deloitte Consulting.

Demirer, M., J. Horton, N. Immorlica, B. Lucier, and P. Shahidi (2025). The economic impacts of generative AI on the structure of work. *Working Paper*.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). GPTs are GPTs: Labor Market Impact Potential of LLMs. *Science 384*(6702), 1306–1308.

Eurostat (2025). Use of Artificial Intelligence in Enterprises. Technical report, European Commission.

Felten, E., M. Raj, and R. Seamans (2021). Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and its Potential Uses. *Strategic Management*

Journal 42(12), 2195–2217.

Felten, E. W., M. Raj, and R. Seamans (2018). A Method to Link Advances in Artificial Intelligence to Occupational Abilities. *AEA Papers and Proceedings 108*, 54–57.

Felten, E. W., M. Raj, and R. Seamans (2023). Occupational Heterogeneity in Exposure to Generative AI. *Available at SSRN* (4414065).

Fradkin, A. (2025). Demand for LLMs: Descriptive Evidence on Substitution, Market Expansion, and Multihoming. *arXiv preprint* (2504.15440).

Hampole, M., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2025). Artificial Intelligence and the Labor Market. *NBER Working Paper* (33509).

Handa, K., T. Eloundou, A. Thakkar, R. Mansfield, and D. Rock (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. *arXiv preprint* (2503.04761).

Humlum, A. and E. Vestergaard (2025). The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities Among Workers. *Proceedings of the National Academy of Sciences 122*(1), e2414972121.

Kong Inc. (2024). API Impact Report 2024: AI Adoption and Innovation Challenges. Technical report, Kong.

McKinsey (2024). The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value. Technical report, McKinsey & Company.

Menlo Ventures (2025). 2025 Mid-Year LLM Market Update: Foundation Model Landscape + Economics. Technical report, Menlo Ventures.

Nagle, F. and D. Yue (2025). The Latent Role of Open Models in the AI Economy. *Available at SSRN* (5767103).

Noy, S. and W. Zhang (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science 381*(6654), 187–192.

OpenAI (2025). The state of enterprise ai. Technical report, OpenAI, San Francisco, CA. OpenAI 2025 Report.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The Impact of AI on Developer Productivity: Evidence from Github Copilot. *arXiv preprint* (2302.06590).

Shao, Y., H. Zope, Y. Jiang, J. Pei, D. Nguyen, E. Brynjolfsson, and D. Yang (2025). Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce. *arXiv preprint* (2506.06576).

Stanford Institute for Human-Centered Artificial Intelligence (2025). The 2025 AI Index Report. Technical report, Stanford University.

The Emerging Market for Intelligence: Supply, Demand, and Pricing of LLMs

Mert Demirer Andrey Fradkin Nadav Tadelis Sida Peng

Appendix

A Additional Tables

Table OA-1: Price Trends Regression Results

	Log Price per Million Prompt Tokens		
	(1)	(2)	(3)
Time Trend	-0.002***	-0.0003***	-8.92×10^{-6}
	(0.0003)	(0.0001)	(9.59×10^{-5})
Time Trend × Open Source			-0.0006***
			(0.0002)
R^2	0.026	0.970	0.970
Observations	98,959	98,959	98,882
Model fixed effects		✓	✓

Notes: This table reports regressions of log prompt token price (per million tokens) on a linear time trend. Column (1) includes only the time trend. Column (2) adds model fixed effects. Column (3) further includes an interaction between the time trend and an indicator for open-source models, allowing price trends to differ between open- and closed-source models. Standard errors, reported in parentheses, are clustered at the model level.

Table OA-2: Inference Providers and Token Shares (December 2025)

Provider	Share (%)	Distinct Creators	Open Source	Closed Source
xai	32.22	1	0	7
google-vertex	20.14	4	1	13
openai	7.37	2	0	15
google-ai-studio	5.62	2	0	13
anthropic	4.84	2	0	11
deepinfra	4.35	11	24	0
novita	4.03	10	24	0
chutes	3.30	7	23	0
minimax	2.96	1	4	0
stealth	2.16	1	4	0
z-ai	1.45	1	7	0
amazon-bedrock	1.37	4	2	14
nebius	0.92	8	18	0
atlas-cloud	0.91	10	19	0
groq	0.81	5	12	0
crusoe	0.73	5	6	0
siliconflow	0.69	7	19	0
mistral	0.62	2	7	12
gmicloud	0.56	8	15	0
deepseek	0.50	1	3	0
parasail	0.48	12	23	0
fireworks	0.48	8	17	0
azure	0.46	5	3	11
together	0.40	11	32	0
moonshotai	0.38	1	4	0
alibaba	0.27	2	22	0
wandb	0.24	5	8	0
cerebras	0.20	4	7	0
baseten	0.19	6	10	0
hyperbolic	0.17	7	20	1
nvidia	0.15	1	2	0
ncompass	0.13	4	6	0
streamlake	0.11	1	1	0
phala	0.11	4	6	0
venice	0.10	5	8	0
friendli	0.09	5	11	0
nextbit	0.09	12	18	0
modelrun	0.07	2	2	0
perplexity	0.06	1	0	6
sambanova	0.05	5	10	0
clarifai	0.04	2	3	0
mancer	0.04	7	9	0
amazon-nova	0.03	1	1	0
cloudflare	0.02	7	13	0
liquid	0.02	1	2	0
avian	0.02	2	2	0
byteplus	0.02	3	4	0
morph	0.01	1	2	0
aion-labs	0.01	1	3	0
infermatic	0.01	3	3	0
open-inference	0.01	5	6	0
meta	0.01	1	4	0
cohere	0.00	1	1	3
inception	0.00	1	2	0
arcee-ai	0.00	1	1	0
relace	0.00	1	1	0
ai21	0.00	1	2	0
switchpoint	0.00	1	1	0
inflection	0.00	1	2	0
featherless	0.00	2	2	0
cirrascale	0.00	1	1	0

Table OA-3: Overview of Benchmark Suite Used by Artificial Analysis

Benchmark	Description
MMLU-Pro	Advanced version of the Multi-Task Language Understanding benchmark with 12,032 10-option multiple-choice questions across science, law, economics, health, and other domains. Evaluates broad reasoning and knowledge.
HLE	Humanities Last Exam is 2,684 challenging open-answer questions across math, humanities, and natural sciences. Designed to test models on very challenging academic tasks.
AA-LCR	Artificial Analysis Long Context Reasoning evaluates reasoning over long contexts (up to 100k tokens) using 100 hard open-answer text-based questions from documents such as reports, consultations, and legal texts.
GPQA Diamond	Scientific reasoning benchmark with 198 graduate-level 4-option multiple-choice questions across biology, physics, and chemistry. Focuses on "Google-proof" knowledge.
AIME 2025	30 questions from the 2025 American Invitational Mathematics Examination with integer answers.
AIME	30 questions from previous American Invitational Mathematics Examinations with integer answers.
MATH-500	500 open-answer mathematics problems assessing high-level symbolic reasoning and competition-style problem solving.
IFBench	Instruction-following benchmark of 294 questions. Tests precise compliance with instructions in a single turn such as counting, formatting, and manipulation.
SciCode	Scientific code generation benchmark of 338 python programming tasks.
LiveCodeBench	Coding benchmark of 315 tasks using python from LeetCode, AtCoder, and Codeforces.
Artificial Analysis Indices	
Intelligence Index	Composite measure aggregating eight constituent benchmarks: MMLU-Pro, HLE, GPQA Diamond, AIME 2025, SciCode, LiveCodeBench, IFBench, and AA-LCR.
Math Index	Reflects math problem solving using AIME 2025 benchmark.
Coding Index	Composite of LiveCodeBench and SciCode benchmarks to reflect programming ability.

Table OA-4: Model Launch Impact: Total Tokens of Existing Models

Launch Event	Model	Before (B)	After (B)	Change (%)
Gemini 2 0 Flash 001	**Gemini 2 0 Flash 001**	0.00	53.22	New
	Claude 3 5 Sonnet	274.82	297.78	+8.4
	Gemini Flash 1 5	39.82	70.60	+77.3
	Gpt 4o Mini	25.66	25.77	+0.4
Claude 3 7 Sonnet	**Claude 3 7 Sonnet**	0.00	213.65	New
	Gemini 2 0 Flash 001	295.24	320.82	+8.7
	Claude 3 5 Sonnet	242.86	108.92	-55.2
	Gemini Flash 1 5	61.68	67.66	+9.7
Gemini 2 5 Pro	**Gemini 2 5 Pro**	136.37	161.45	+18.4
	Claude 3 7 Sonnet	367.52	345.48	-6.0
	Gpt 4o Mini	113.54	326.28	+187.4
	Gemini 2 0 Flash 001	286.24	247.96	-13.4
Gemini 2 5 Flash	**Gemini 2 5 Flash**	0.00	90.65	New
	Claude 3 7 Sonnet	355.11	376.72	+6.1
	Gemini 2 0 Flash 001	260.33	202.65	-22.2
	Gemini 2 5 Pro	193.39	201.23	+4.1
Claude 4 Sonnet	**Claude 4 Sonnet**	0.00	240.01	New
	Gpt 4o Mini	453.06	497.48	+9.8
	Claude 3 7 Sonnet	395.17	312.28	-21.0
	Gemini 2 5 Flash	185.27	223.24	+20.5
Gemini 2 5 Flash Lite	**Gemini 2 5 Flash Lite**	0.00	47.69	New
	Claude 4 Sonnet	284.49	365.70	+28.5
	Deepseek Chat V3	238.29	263.15	+10.4
	Gemini 2 5 Flash	237.44	259.47	+9.3
Grok Code Fast 1	**Grok Code Fast 1**	0.00	536.52	New
	Claude 4 Sonnet	527.38	549.41	+4.2
	Gemini 2 5 Flash	251.43	470.08	+87.0
	Deepseek Chat V3	262.96	218.55	-16.9
Grok 4 Fast	**Grok 4 Fast**	0.00	635.08	New
	Grok Code Fast 1	1181.38	1029.39	-12.9
	Claude 4 Sonnet	604.98	548.77	-9.3
	Gemini 2 5 Flash	325.94	349.62	+7.3
Claude 4 5 Sonnet	**Claude 4 5 Sonnet**	0.00	298.52	New
	Grok Code Fast 1	1070.80	1057.91	-1.2
	Grok 4 Fast	1015.38	738.80	-27.2
	Claude 4 Sonnet	579.29	315.83	-45.5
Gemini 3 Pro	**Gemini 3 Pro**	0.00	191.92	New
	Grok Code Fast 1	1452.95	1204.38	-17.1
	Claude 4 5 Sonnet	605.17	495.98	-18.0
	Gemini 2 5 Flash	444.19	448.66	+1.0

Note: Token usage in billions.

Table OA-5: Price Elasticity Regressions - No Provider Performance Covariates

	Log(Daily Tokens)		
	(1)	(2)	(3)
Log(Price)	-0.48***	-1.07***	-1.02***
	(0.15)	(0.23)	(0.13)
Observations	48,362	48,362	48,362
R^2	0.11	0.67	0.93
Within R^2	0.07	0.18	0.05
Date fixed effects	✓	✓	
Model fixed effects		✓	
Date × Model fixed effects			✓
Model × Provider fixed effects			✓

Notes: This table reports regressions of log daily token usage on log prompt token price. Each column shows a different fixed-effects specification. Column (1) includes date fixed effects. Column (2) adds model fixed effects. Column (3) includes date × model fixed effects and model × provider fixed effects. Standard errors are reported in parentheses.

Table OA-6: Correlation of Prices with Provider Performance

	Log(Price)		
	(1)	(2)	(3)
Log(Throughput + 1)	0.08*		
	(0.04)		
Missing Throughput	1.02***		
	(0.23)		
Log(Latency + 1)		0.01	
		(0.08)	
Missing Latency		0.72***	
		(0.13)	
Log(Context Length + 1)			-0.13
			(0.12)
Observations	48,362	48,362	48,362
R^2	0.84	0.84	0.79
Within R^2	0.22	0.21	0.01
Date fixed effects	✓	✓	✓
Model fixed effects	✓	✓	✓

Notes: This table reports regressions of log model prices on provider performance characteristics. The dependent variable in all columns is log price per million prompt tokens. Column (1) relates prices to log(throughput + 1) and an indicator for missing throughput. Column (2) includes log(latency + 1) and a missing-latency indicator. Column (3) includes log(context length + 1). All specifications include date fixed effects and model fixed effects, and standard errors are reported in parentheses.

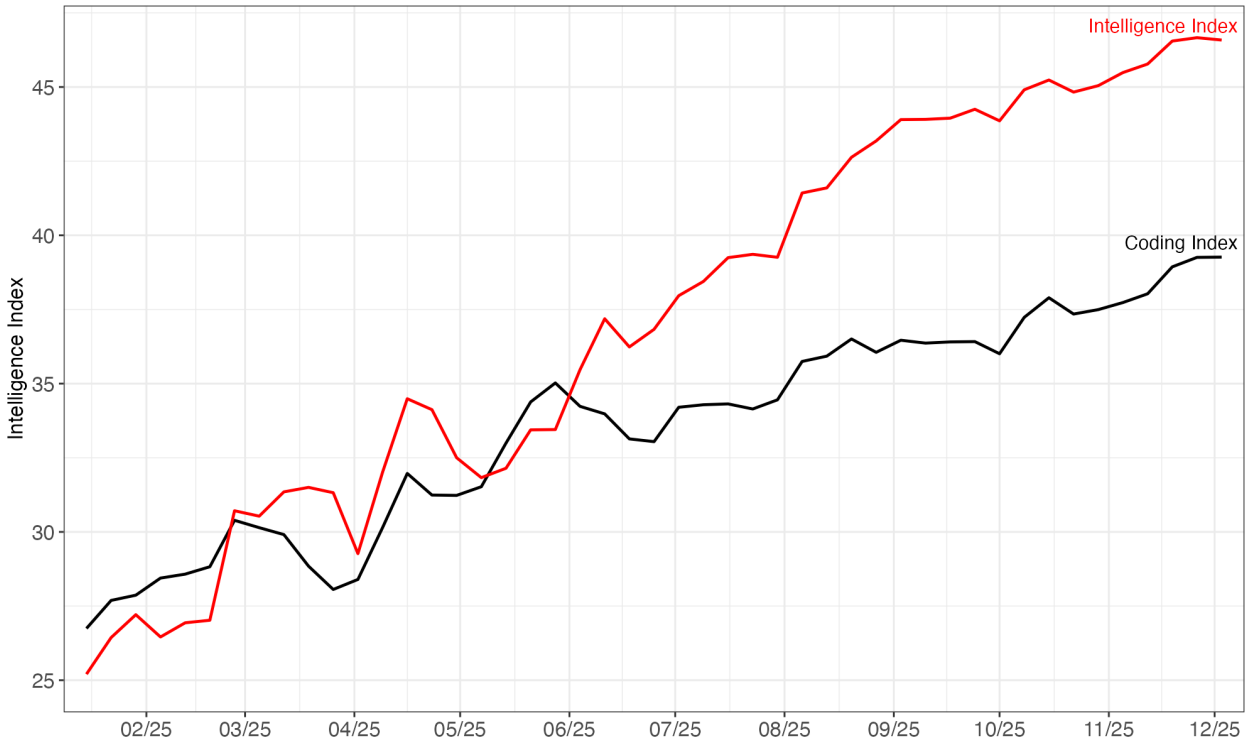
Table OA-7: Model Naming Convention Examples

model_name5	model_name4	model_name3	model_name2	model_name1
claude-4.5-sonnet-20250929	claude-4.5-sonnet-20250929	claude-4.5-sonnet	claude-4.5-sonnet	claude-4.5
deepseek-chat-v3-0324	deepseek-chat-v3-0324	deepseek-chat-v3	deepseek-chat-v3	deepseek-chat
gemini-2.5-flash	gemini-2.5-flash	gemini-2.5-flash	gemini-2.5-flash	gemini-2.5
gemini-3-pro-preview-20251117	gemini-3-pro-20251117	gemini-3-pro	gemini-3-pro	gemini-3
gpt-5-mini-2025-08-07	gpt-5-mini-2025-08-07	gpt-5-mini	gpt-5-mini	gpt-5
llama-3.3-70b-instruct	llama-3.3-70b-instruct	llama-3.3	llama-3.3	llama-3.3
mercury-coder-small-beta	mercury-coder-small	mercury-coder-small	mercury-coder	mercury

Notes: This table illustrates the systematic naming convention used in the paper. Each row shows a model from a different creator, progressing from the raw name as observed in the data to the most aggregated level. Raw Name (model_name) is the original model name. model_name5 removes free/paid distinctions for technically identical models. model_name4 removes beta/preview release references. model_name3 removes date information. model_name2 distinguishes reasoning vs. non-reasoning and other major variants. model_name1 represents the most general model family definition.

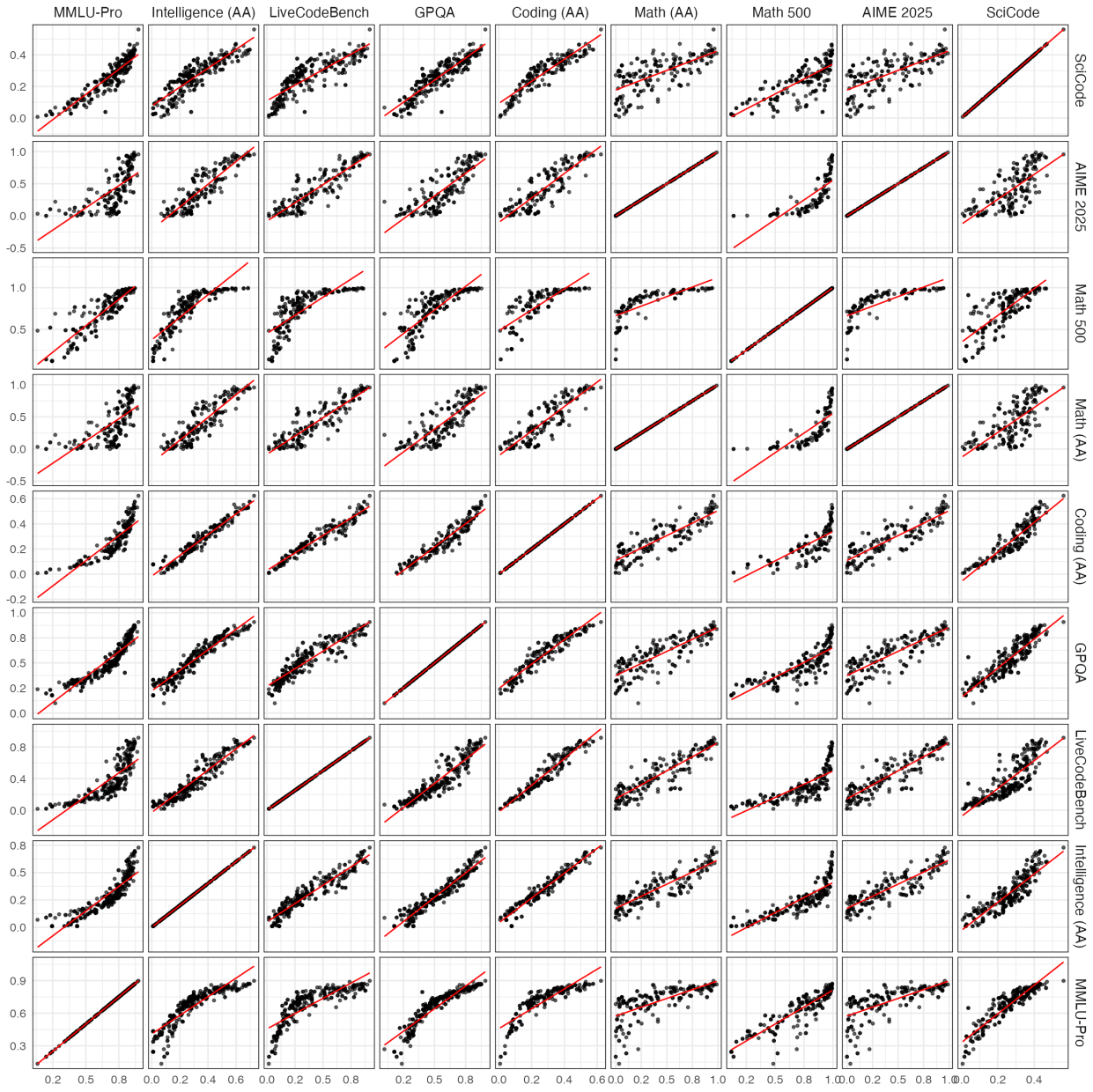
B Additional Figures

Figure OA-1: Average Intelligence of API Calls



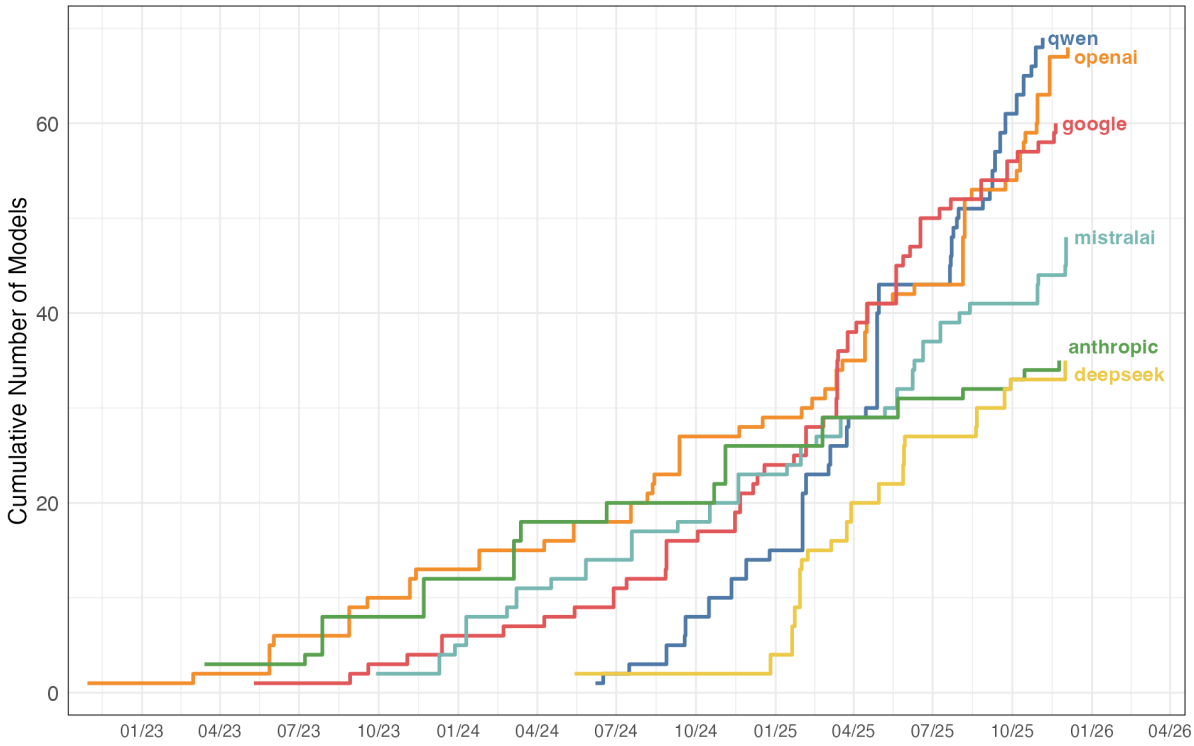
Notes: This figure shows the average intelligence of API calls over time. The blue line plots the mean Intelligence Index per token, while the red line plots the mean Coding Index per token. The x-axis is calendar time (January through December 2025), and the y-axis reports the respective index values.

Figure OA-2: Correlation Between Different LLM Benchmarks



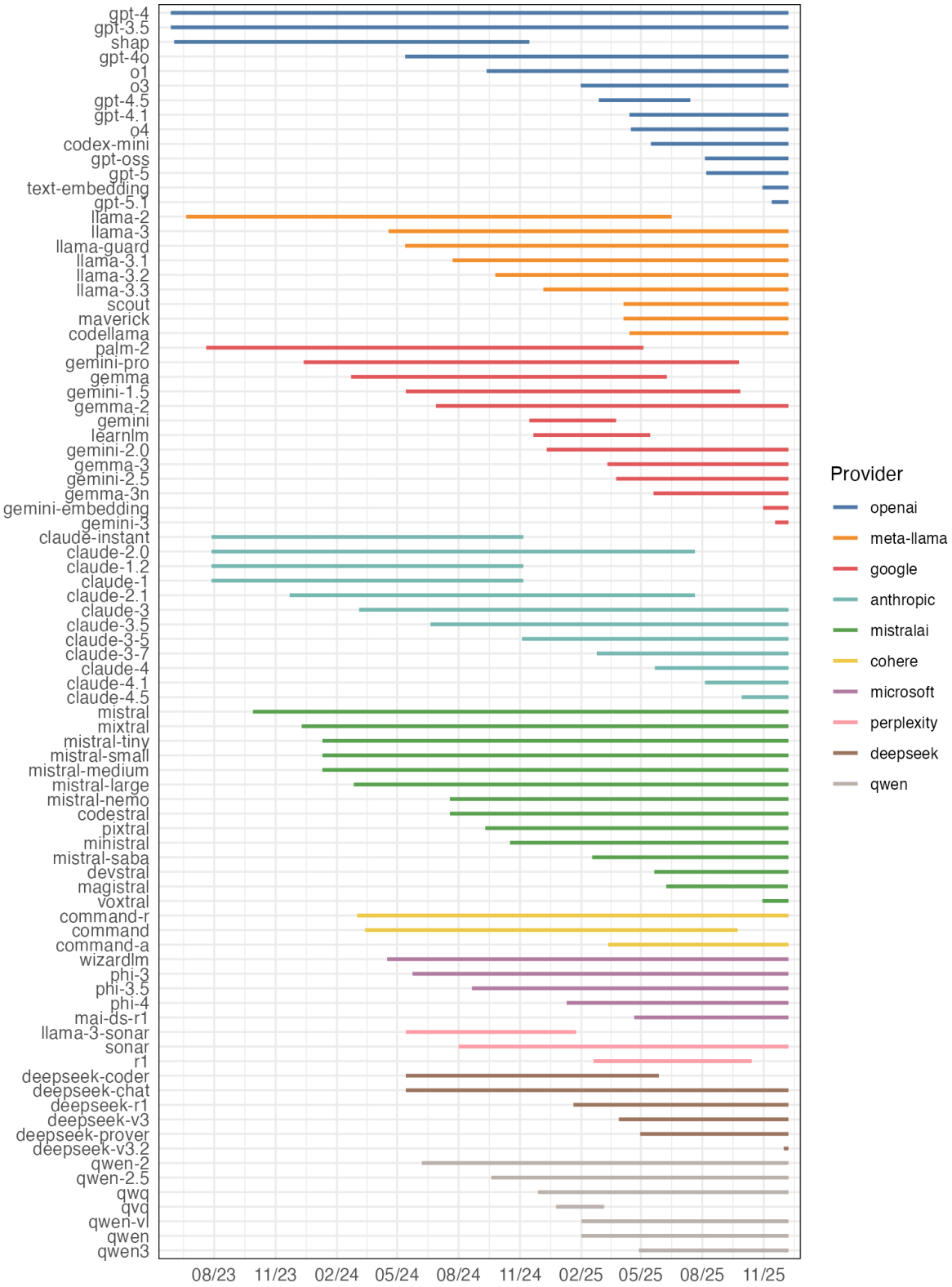
Notes: The figure reports linear relationships across nine large language model benchmarks. The included benchmarks include MMLU-Pro (Massive Multitask Language Understanding Pro) which is a test with questions across dozens of academic subjects designed to measure broad reasoning and subject-matter expertise; Intelligence (AA) is the Intelligence Index from Artificial Analysis; LiveCodeBench is a coding benchmark that evaluates models on writing and debugging code; GPQA is the Graduate-Level Google-Proof Q&A which is a set of challenging science and reasoning questions that cannot be solved by simple web search; Coding (AA) is the Coding Index from Artificial Analysis; Math (AA) is the math index from Artificial Analysis; Math 500 is a set of 500 challenging math problems from OpenAI's math benchmark test; AIME 2025 is based on problems from the 2025 American Invitational Mathematics Examination; SciCode is a benchmark designed to test writing code for solving scientific research problems.

Figure OA-3: Cumulative Models Created Over Time by Creator



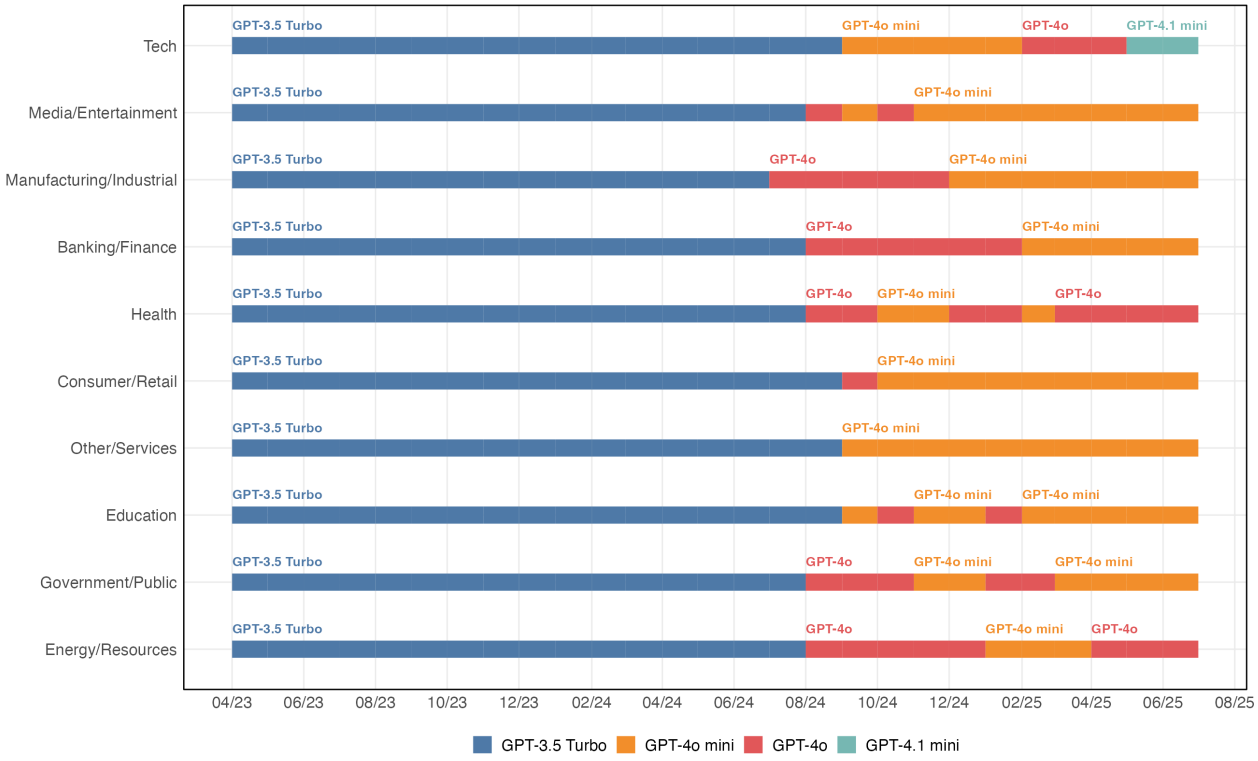
Notes: Cumulative number of models created over time by major developers. The x-axis shows observations from December 2022 through December 2025, and the y-axis shows the cumulative number of models. Separate lines are plotted for Google, OpenAI, Anthropic, Mistral, Deepseek, and Qwen.

Figure OA-4: Availability of Model Families Over Time for the Top 10 Creators



Notes: This figure shows the availability of model families developed by the top 10 providers. Each bar represents the time period during which a model family was available.

Figure OA-5: Top Model of Each Industry for Firms using Microsoft Azure

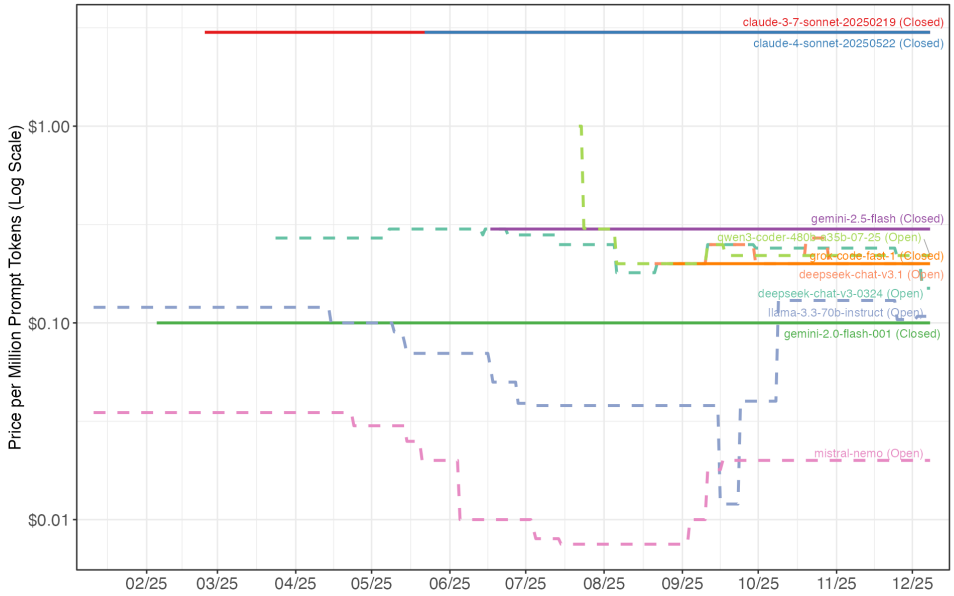


Notes: This figure shows the market-leading model for firms in each industry that use LLMs hosted on Microsoft Azure Foundry from April 2023 through July 2025.

Figure OA-6: Availability of Models Over Time for the Top 10 Creators

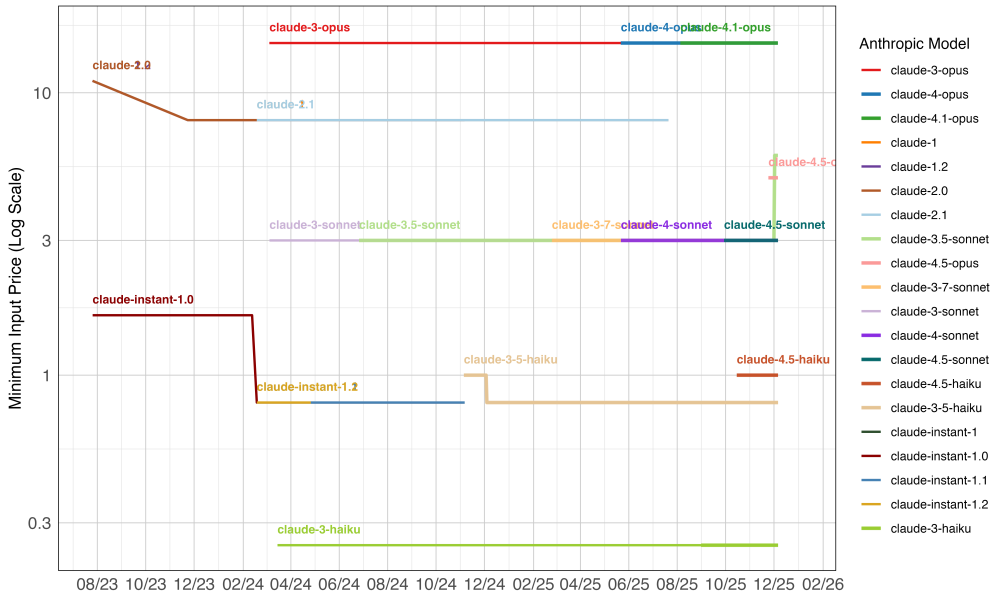
Notes: This figure shows the availability of individual models developed by the top 10 creators determined by number of models. Each bar represents the time period during which a model was available.

Figure OA-7: Prompt Token Price Evolution - Top Five Open and Closed-Source Models



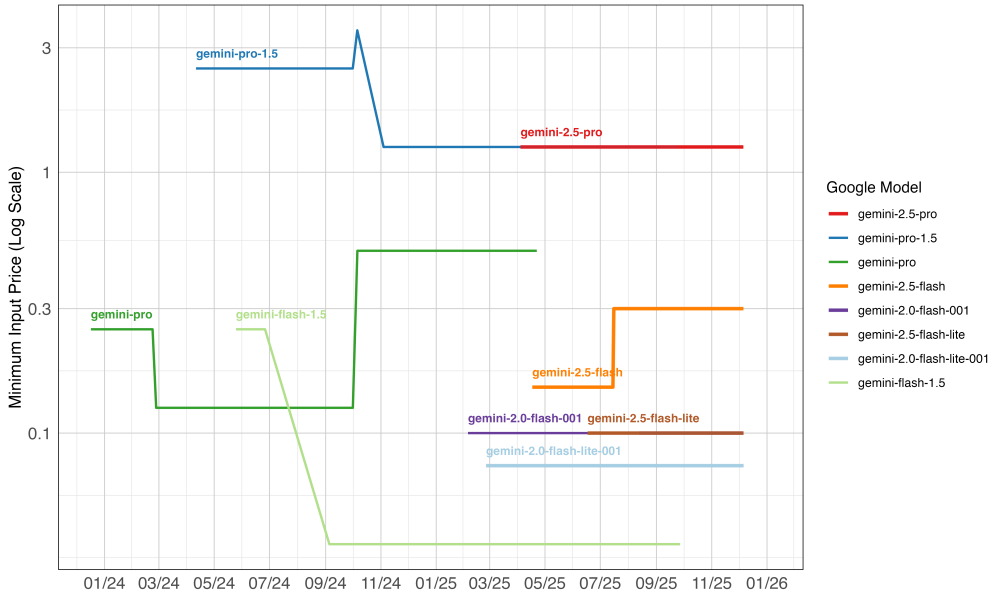
Notes: Top models are categorized by total token usage (solid lines indicate closed-source, dashed lines indicate open-source). Labels show model names with source type at the end of each line.

Figure OA-8: Price Trajectories for Anthropic Models Over Time



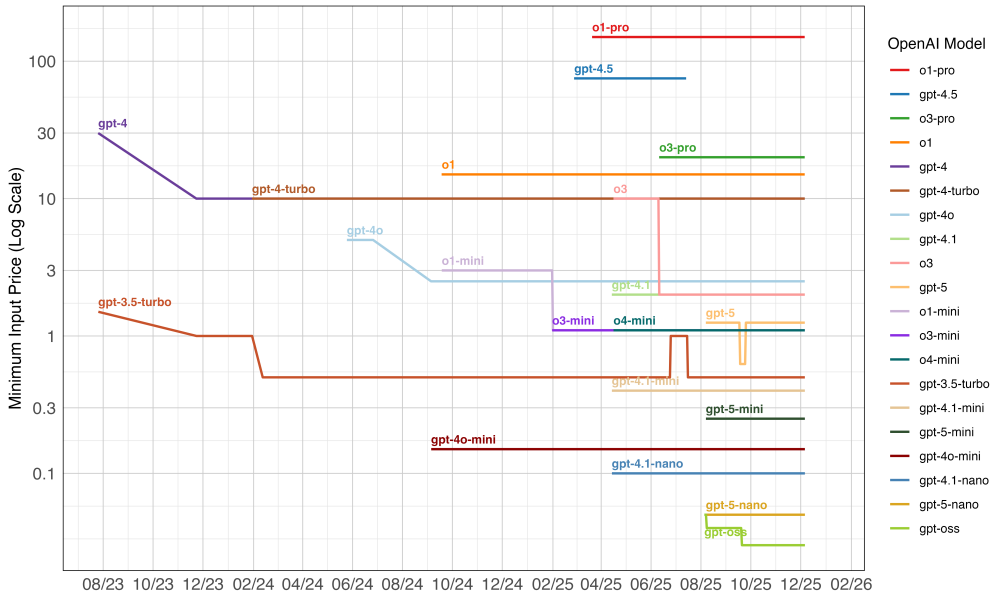
Notes: Each line shows the minimum prompt price for a given Anthropic model variant over time. The y-axis is on a log scale. The legend is ordered by final observed price.

Figure OA-9: Price Trajectories for Google Models Over Time



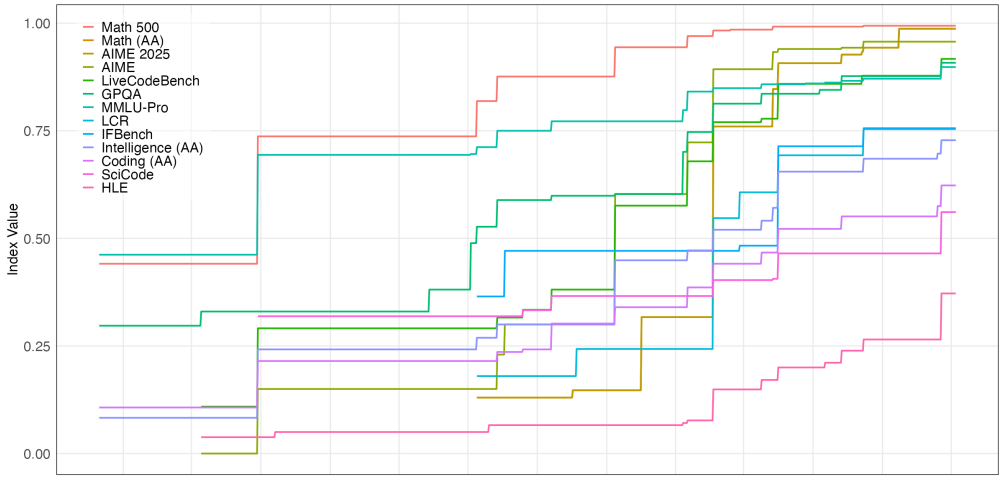
Notes: Each line shows the minimum prompt price for a given Google model variant over time. The y-axis is on a log scale. The legend is ordered by final observed price.

Figure OA-10: Price Trajectories for OpenAI Models Over Time



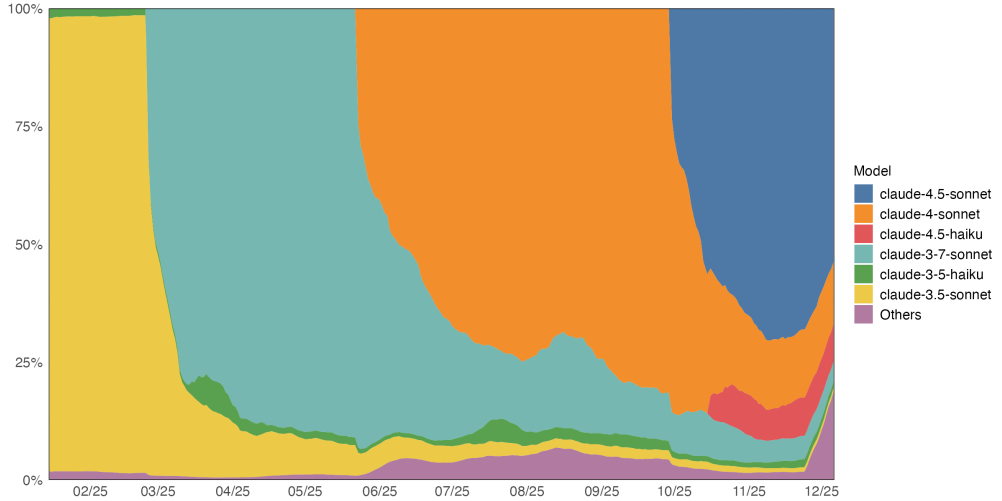
Notes: Each line shows the minimum prompt price for a given OpenAI model variant over time. The y-axis is on a log scale. The legend is ordered by final observed price.

Figure OA-11: Daily Maximum Benchmark Index Values Over Time



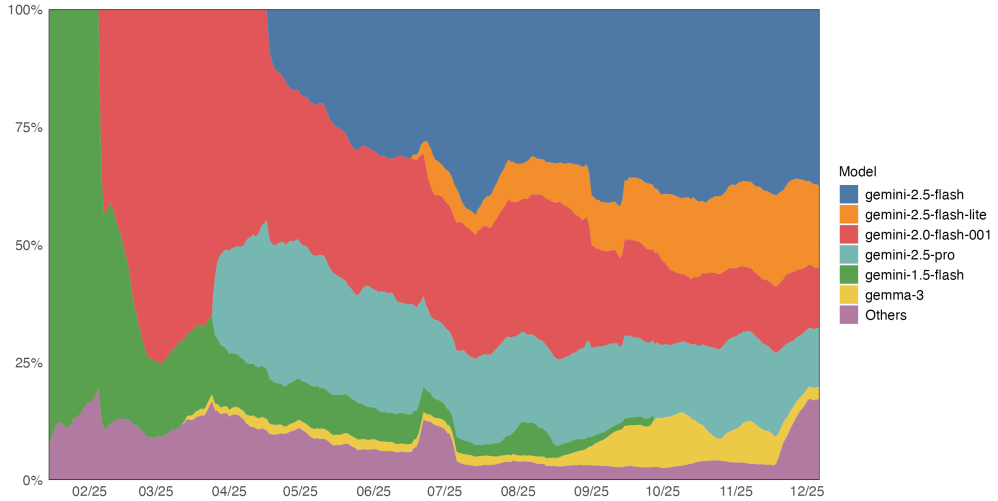
Notes: The figure plots the daily maximum values across a set of benchmark indices, including math, coding, reasoning, and composite indices. See OA-3 for an overview of all included benchmarks. line represents a benchmark, normalized between 0 and 1. Step increases indicate the introduction of new models achieving higher benchmark scores.

Figure OA-12: Model Shares for Anthropic



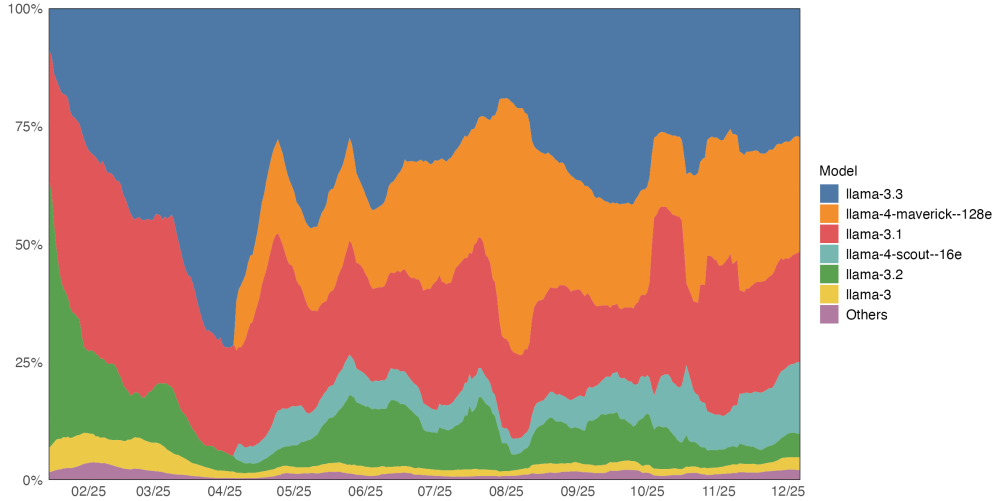
Notes: Share of Anthropic's usage by model (14-day rolling average) from November 2024 to December 2025. Top six models are displayed: Claude 3.5, 3.7, and 4 series dominate usage, with rapid shifts following each new release

Figure OA-13: Model Shares for Google



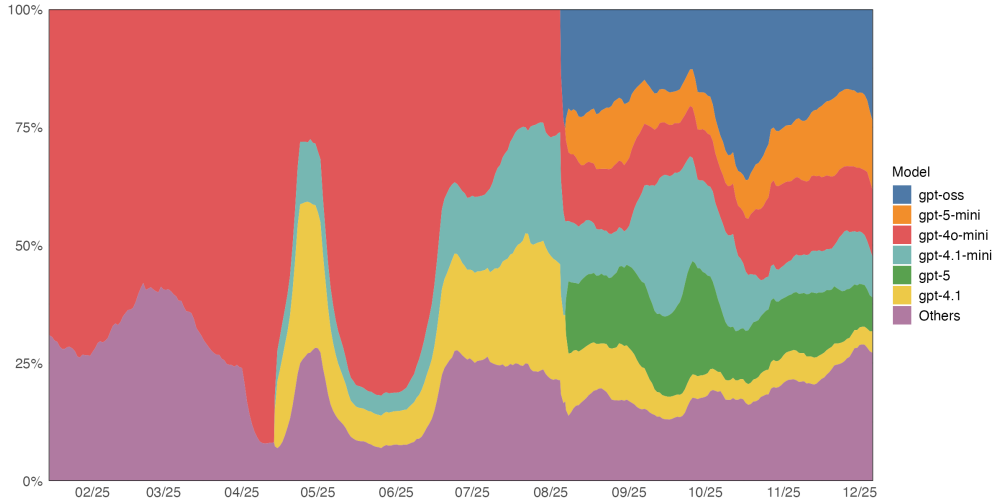
Notes: Share of Google's usage by model (14-day rolling average). Usage is concentrated in Gemini 2.0 and 2.5 Flash/Pro variants, which gain substantial share after release

Figure OA-14: Model Shares for Meta-LLaMA



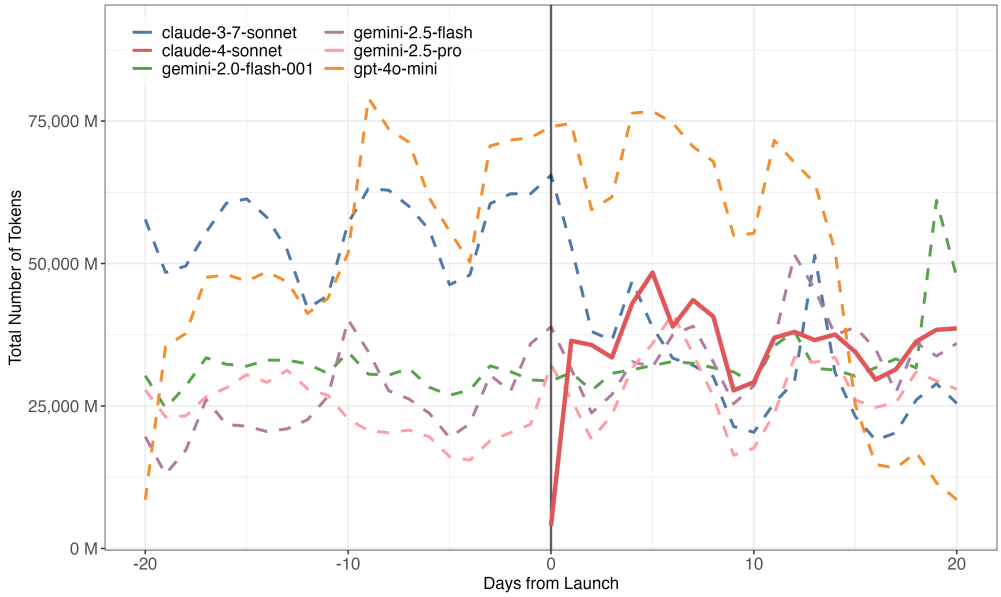
Notes: Share of Meta-LLaMA's usage by model (14-day rolling average). LLaMA 3.x models hold steady shares, with LLaMA 4 Maverick and Scout appearing later in the sample

Figure OA-15: Model Shares for OpenAI



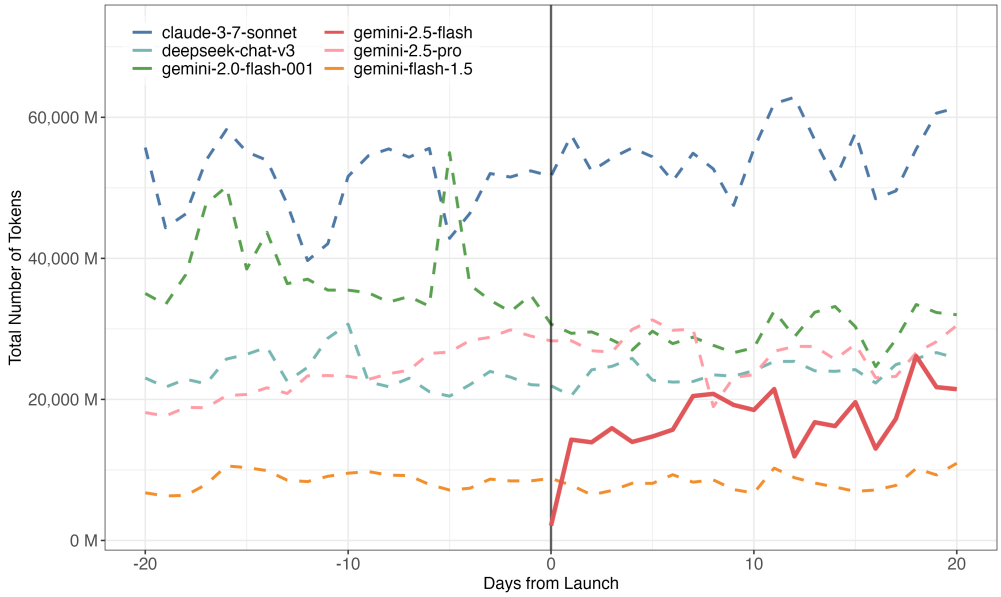
Notes: Share of OpenAI's usage by model (14-day rolling average). GPT-4o Mini dominates early in the sample, later supplemented by GPT-4.1 and GPT-5 variants)

Figure OA-16: Usage for Select Models Following Claude 4 Sonnet Release



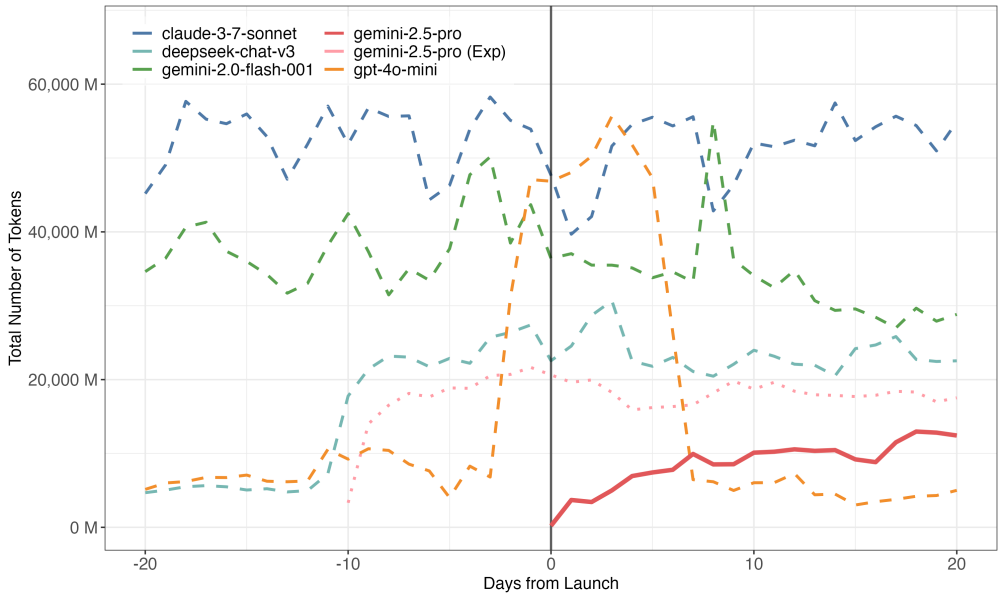
Notes: Daily token usage for selected models around the Claude 4 Sonnet launch. The x-axis is days from launch, with a vertical black line at day 0 (window $\approx$ -20 to +20 days). Series include `claude-4-sonnet` (focal), `claude-3.7-sonnet`, and other leading contemporaries. y-axis units are millions of tokens.

Figure OA-17: Usage for Select Models Following Gemini 2.5 Flash Release



Notes: Daily token usage for selected models around the Gemini 2.5 Flash launch. The x-axis is days from launch with a vertical black line at day 0 (window $\approx$ -20 to +20 days); the y-axis reports tokens. Series include `gemini-2.5-flash` (focal) and contemporaneous leading models.

Figure OA-18: Usage for Select Models Following Gemini 2.5 Pro Release



Notes: Daily token usage for selected models around the Gemini 2.5 Pro launch. The x-axis is days from launch with a vertical black line at day 0 (window $\approx$ -20 to +20 days); the y-axis reports tokens. Series include `gemini-2.5-pro` (focal) and contemporaneous leading models (e.g., `deepseek-chat-v3`, `claude-3.7-sonnet`).

C Data Appendix

This section describes the datasets.

C.1 OpenRouter Data

We aggregate publicly available information on OpenRouter's website. OpenRouter is a marketplace for LLMs and a popular API gateway for app developers. They provide access to hundreds of LLMs from many creators such as OpenAI, Anthropic, Google, Meta, and Deepseek. In practice, the OpenRouter platform sits between app developers and LLM inference providers, routing traffic to selected providers while aggregating pricing and usage analytics. OpenRouter publishes these pricing and usage analytics for each provider as well as a number of other model-related attributes.

OpenRouter provides a model catalog with an individual page per model. A model is defined at the model-variant level; for instance, some LLMs have a free variant in addition to a standard variant, or some might have a beta variant in addition to a standard variant. In these cases, there will be two model pages, and we treat each as a separate model. An LLM could potentially have more than two pages (e.g. free, beta, and standard variants), or just one page (standard variant only).

C.1.1 Model and Provider Data

On each model page, OpenRouter lists its providers; providers are platforms that host LLMs and provide the computing infrastructure needed to deploy them. Some models have a single provider, while other open-source models can have many providers. Providers of open-source models include independent hosts such as Chutes, DeepInfra, Novita AI, and Nebius, among others. Some providers, such as Anthropic and OpenAI, host only models they themselves develop. OpenRouter reports the current prices per million tokens separately for prompt and completion tokens. When relevant, pages also display additional prices, such as for image or audio tokens.

OpenRouter also publishes reliability and capacity statistics at the provider level on the model page: recent latency, throughput, uptime, supported context window, and maximum completion (measured in completion tokens). It also includes information on the provider's data policy.

Beyond provider-level information, each model page reports recent activity as daily tokens used over the last 90 days for both prompts and completion, "Top public apps this week" which are the names and urls for the top 20 apps ranked by total weekly token usage of a given model, and a number of headline specs on each model including the model creator, the creation date, model capability description, model group, and the context window. There are

also headline prompt and completion prices which are the minimum prices available from the available providers. The app-level usage data is only for apps that opt to have their usage tracked and reported by OpenRouter; they receive a discount for doing so.

The following summarizes important information found on OpenRouter's model pages.

- **Model variant** — The specific version of a model (e.g. standard, beta, free, thinking, extended). Each variant has its own page and may differ in features or pricing
- **Provider** — The platform that hosts and serves the model endpoint (e.g., OpenAI, Anthropic, DeepInfra). A model can be offered by one or many providers
- **Prompt price** — Price per million prompt tokens counted on the request side
- **Completion price** — Price per million completion tokens generated by the model in its response
- **Total context** — The maximum amount of information (measured in tokens) a model can keep in its working memory to refer to for a given conversation or task
- **Max output** — The maximum length (measured in tokens) of a single response returned by the model
- **Group** — A label that ties model variants to a model family (e.g. GPT, Claude, Gemini)
- **Creation date** — The date the model variant was created
- **Uptime** — The share of requests succeeding based on traffic routed through the platform's endpoints for the model over the last 3 days
- **Latency** — Response time to first token
- **Throughput** — Average number of completion tokens per second over the last 30 minutes
- **Recent activity** — Total model usage in prompt tokens and completion tokens per day on OpenRouter, aggregated across all providers and reported for the last 90 days
- **Top public apps** — Top 20 apps using the model who have opted to have their usage published by OpenRouter. Top apps are determined by total token usage over the last 7 days.

C.1.2 Leaderboard Data

In addition to the model pages, OpenRouter publishes several leaderboards determined by total token usage routed through its API gateway. Rankings are reported at daily, weekly, and monthly frequencies for the top models overall. At the weekly level, OpenRouter reports top models across a set of usecases including programming, roleplay, marketing, marketing/seo, technology, science, translation, legal, finance, health, trivia, and academia. Weekly rankings are also posted for total token usage by model creator and the top 20 apps that use the most tokens across all models. Lastly, OpenRouter publishes a page for each provider with the daily total tokens routed through that provider for the top 10 models.

C.1.3 Data Methodology

OpenRouter's model catalog reports current pricing information for all models that are active (they offer current API access to). OpenRouter's model catalog also includes some deprecated models without pricing information or current usage data. We focus on the active models. Our core OpenRouter dataset consists of information aggregated from webscrapes of OpenRouter's individual model pages for all active models. From April 11, 2025 to December 8, 2025, we scraped each model page as of 6:00 AM UTC. As stated previously, OpenRouter identifies models at the variant level and tracks them over time with a model variant permaslug; this permaslug remains constant even when the model is updated or pricing changes. We use this permaslug as a panel identifier.

We compile a daily model-provider pricing panel from webscraped pages of each active model on OpenRouter from April 11, 2025 through December 8, 2025. Specifically, we focus on models for which OpenRouter reports pricing data, as pricing data are removed from deprecated models. Recent activity on each model page is reported for the past 90 days so we have usage data from January 11, 2025. Additionally, we use the internet archive to get model information as far back as November 2023 although these records are not consistently captured each day. We construct a panel on provider-model pricing.

For the leaderboard data, we only began scraping these pages directly from OpenRouter on August 14, 2025, this data is reported for the past three months so we have consistent data from May 14, 2025. We supplement this data with pages pulled from the internet archive. We focus on the usecase rankings and compile a dataset of the top 10 models per usecase per week. Combining daily scrapes with internet archive data, we see weekly category rankings from November 2023 through December 2025.

The following table provides an overview of what raw data is available and its level of consistency. We are able to build a complete panel with weekly usage and pricing for each active model from January 1, 2025 through December 8, 2025, which serves as our core

dataset.

Table OA-8: Overview of Data Availability

Data	Dates Available	Consistency
Daily model-level token usage	Jan 2025 - Dec 2025	Complete panel starting Jan 11, 2025
Weekly model-level token usage	Nov 2024 - Dec 2025	49 percent of days observed until complete panel starting Aug 14, 2025 .
Daily category rankings	Nov 2024 - Dec 2025	49 percent of days observed until complete panel starting Aug 14, 2025.
Top 20 apps by weekly usage by model	Apr 2025 - Dec 2025	Complete panel starting April 11, 2025
Headline model–level prices and specs	Nov 2023 - Dec 2025	31 percent of days until complete panel starting April 11, 2025.
Model–provider pricing	Nov 2023 - Dec 2025	6 percent of days until complete panel starting April 11, 2025.
Provider-level usage	May 2025 - Dec 2025	Complete panel for top 10 models for each provider starting May 17, 2025
Provider-level reliability specs	Apr 2025 - Dec 2025	Daily panel starting April 11, 2025 with high missingness

D OpenRouter Data Processing

D.1 Model Naming

Defining what constitutes a "model" is non-trivial. In practice, models often exist in multiple variants, including versions launched on specific dates, beta or preview releases, and free versus paid offerings. In our dataset, model names are recorded at the level presented to the user during selection. This results in a highly granular set of identifiers that is not always appropriate for analysis.

To address these issues, we develop a systematic naming convention that assigns models to increasingly aggregated levels. We first remove any information on whether the model is free or paid, since these offerings are technically identical and differ only in pricing; we denote this cleaned name as modelname5. Next, we strip references to beta or preview releases, yielding modelname4. We then remove the date information, producing modelname3. Finally, we manually classify models into two broader categories: modelname2, which reflects distinctions such as reasoning versus non-reasoning variants, and modelname1, which corresponds to the most general definition of a model family.

D.2 Merging with Artificial Analysis

We merge our data with benchmark scores from Artificial Analysis in order to obtain performance measures for the models. Because the naming conventions used by OpenRouter and Artificial Analysis do not align, we manually matched models across the two sources. The matching was performed at the modelname3 level, which corresponds to the most granular model-level information available from Artificial Analysis. This procedure may introduce errors, as the model names reported by Artificial Analysis are often more general than those in OpenRouter, and it is not always clear whether a given OpenRouter model belongs to a particular family in Artificial Analysis. To mitigate this risk, we adopted a conservative approach, including only matches that could be identified with high confidence. We also flagged likely matches and use these only in analyses that are not sensitive to potential matching errors.

D.3 Free Models

Some providers offer free versions of models on OpenRouter, typically subject to usage quotas. Free model usage accounts for approximately 5% of total activity in the OpenRouter data. We exclude free models from all pricing analyses, as they do not reflect the true cost of accessing the model. However, usage of free models is included in analyses of overall usage patterns.

D.4 Secret Models

OpenRouter occasionally displays models under internal-sounding or code-name identifiers, especially when new models are added before full documentation is available. Some appear through the API or community discovery before receiving standard public names. Examples include "Sherlock Alpha" and "Sherlock Think Alpha," which surfaced ahead of formal labeling, as well as more cryptic IDs such as "deepseek/r1-0528" or "mistral-nemotron-super-49b." These labels typically reflect provider-side versioning, experimental releases, or temporary aliases during rollout. As a result, users may encounter models that are assigned obscure code names, even though they correspond to real, accessible LLMs in the OpenRouter ecosystem.

Since the appearance of these models is infrequent and often short-lived, we do not attempt to merge them with the officially released models that appear later. This is a reasonable choice because users themselves may not know the true underlying model or final name at the time these identifiers surface