The Emerging Market for Intelligence: How Firms Buy and Sell AI *

Mert Demirer† Andrey Fradkin† Nadav Tadelis†

June 12, 2026

*We thank Alessandro Bonatti, Seth Benzell, James Brand, Erik Brynjolfsson, Tom Cunningham, Joseph Doyle, Dean Eckles, Garrett Johnson, Daniel Rock, Daniel Waldinger, the editors, and seminar participants at the Stanford Digital Economy Lab, MIT Sloan, and MIT FutureTech. We thank Marilyn Pareboom for exceptional research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the views of Amazon or Microsoft. None of the authors hold a financial interest in Artificial Analysis or OpenRouter. To our knowledge, Amazon and Microsoft have no material financial interest in Artificial Analysis or OpenRouter.

†Mert Demirer is Associate Professor of Economics, MIT Sloan School of Management, Cambridge, Massachusetts. Andrey Fradkin is Associate Professor of Marketing, Boston University Questrom School of Business, Boston, Massachusetts. While writing this paper, he was on leave at Amazon Inc. Nadav Tadelis is Senior Economist in the Office of the Chief Economist, Microsoft Corporation, Seattle, Washington. Their email addresses are mdemirer@mit.edu, fradkin@bu.edu, and nadavtadelis@microsoft.com.

1 Introduction

In the business-to-business market for large language model (LLM) inference, firms pay LLM providers to process their prompts and return output text in response. In late 2023, the price of state-of-the-art artificial intelligence in this market was roughly $30 per million tokens—with each token representing roughly three-quarters of a word of text. In late 2025, the same level of intelligence cost about three cents—a 1,000-fold decline. No other major technology has seen its quality-adjusted price fall so quickly.

Artificial intelligence is widely regarded as a general-purpose technology (Bresnahan and Trajtenberg, 1995; Eloundou et al., 2024), one that, like electricity or the microprocessor, has the potential to reshape production across every sector of the economy. But general-purpose technologies do not instantly diffuse throughout the economy. They require institutions that set prices, match buyers to sellers, and transmit information about quality. The market for LLMs is one such institution, intermediating how AI capabilities are priced, evaluated, and deployed across the economy.

We describe the features of this market and document key empirical patterns in pricing, entry, usage, and market dynamism. Our analysis uses data from OpenRouter, a major intermediary in the LLM ecosystem that connects buyers and sellers. OpenRouter operates a standardized Application Programming Interface (API), a protocol that lets one program send requests to another and receive responses. OpenRouter provides access to hundreds of LLMs from dozens of providers, allowing us to observe pricing, entry, and usage across a broad cross-section of the business AI market from 2023 to 2025.

The supply side of this market has expanded rapidly. Our focus is on models readily available to firms for commercial use through APIs, a set that grew from 270 at the beginning of January 2025 to 668 by December 2025. Behind this proliferation, the number of firms developing these models has nearly doubled, while the number of inference providers—firms that host and serve models—has more than tripled over the same period. (The broader universe of artificial intelligence models is vastly larger still: the Hugging Face platform alone hosts millions of models.)

Competition in the market is intense. The leading model shifts frequently as successive releases displace incumbents at the capability frontier, reflecting rapid vertical innovation. The most intelligent model at any point in time does not win the majority of demand. There is substantial horizontal differentiation: no single model dominates across all use cases, and different models specialize in distinct tasks. These forces create a competitive environment characterized by both rapid obsolescence and specialization.

This proliferation of models and providers recalls similar episodes in economic history.

For example, the US automobile industry had roughly 250 manufacturers by 1910, just a decade after the first commercial vehicles appeared; a shakeout then winnowed the field to a handful of survivors (Klepper, 2002). The internet service provider market of the mid-1990s saw thousands of local ISPs spring up, only to consolidate as scale economies took hold (Greenstein, 2015). We describe the ways in which the LLM market is now in a similar phase. Whether the market for intelligence follows a similar path toward concentration or instead fragments into specialized niches will depend on the strength of scale economies, the pace of capability improvements, and interactions across the industry's vertical layers.

Our focus is on business adoption of LLMs. We do not study consumer use through chat interfaces (Chatterji et al., 2025; Bick et al., 2024), or the broader labor-market (Eloundou et al., 2024; Brynjolfsson et al., 2025; Acemoglu and Restrepo, 2019) and macroeconomic (Acemoglu, 2024) consequences of AI. These are important and rapidly growing areas of inquiry. What LLM markets offer is a unique window into how firms are accessing intelligence as a productive input.[1]

2 How the Market Works

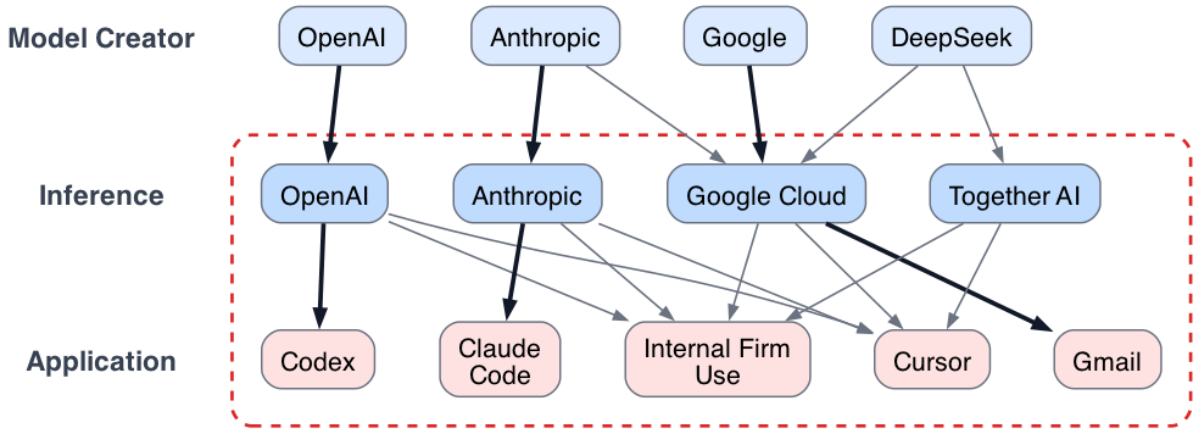
The market for LLM inference is organized as a vertical supply chain with three layers: creators who build models (the model layer), inference providers who host and serve them (the inference layer), and end consumers—firms that either integrate LLM capabilities into their internal workflows or embed them in downstream applications that serve their own customers (the application layer). See Figure 1 for an illustration. Some players occupy multiple positions in the supply chain: all major model creators both train models and serve inference through their own platforms, and several—such as Anthropic and OpenAI—extend further into the application layer by offering their own end-user products, such as coding assistants. In this paper, we focus on the interface between inference providers and customers.

The Supply Chain

Model creators are the firms that design and train LLMs. Training a frontier model requires assembling massive datasets, designing a model architecture, and executing computationally intensive training runs that can cost hundreds of millions of dollars (Epoch AI, 2025). The leading closed-source creators—Anthropic, Google, and OpenAI—keep their model weights proprietary. For example, Google develops models internally and offers them

[1]The numbers and figures in this paper reflect the state of the market as of December 2025, and the text was last updated in April 2026. This market is moving rapidly, and some specifics may be out of date by publication.

Figure 1: Market Structure: Model Layer, Inference Layer, and Application Layer (Illustrative Firms in Each Layer)



Notes: The figure illustrates the vertical structure of the market for LLM inference. The Model layer comprises the frontier labs that train and maintain foundation models. The Inference layer consists of cloud and specialized providers that host models and process API requests. The Application layer includes labs' own products, third-party applications built on top of inference APIs (e.g., Cursor, Perplexity), and internal firm use of LLMs. Bold arrows indicate vertical integration: Google owns Google Cloud and Gmail; OpenAI owns Codex; Anthropic owns Claude Code. Thin arrows indicate market-based supply relationships between independent firms. The dashed box marks the focus of this paper: the interface between inference providers and customers.

directly through its cloud platform. In some cases, closed-source creators distribute models through multiple providers to expand reach—Anthropic, for instance, makes its Claude models available through AWS, Microsoft Azure, and Google Cloud in addition to its own platform.

Other creators, most notably Meta and the Chinese firm DeepSeek, release their model weights under open-source (or open-weight) licenses, allowing anyone to download and run the model. The distinction between open-source and closed-source models is fundamental to the market's competitive structure (Lerner and Tirole, 2005). Closed-source models are available only through the creator or its authorized partners. Open-source models, by contrast, can be hosted by any firm with sufficient computing capacity. They can also be modified: firms routinely fine-tune open-source models for specific tasks or "distill" large models into smaller, faster variants that sacrifice some capability for lower cost.

Inference providers are the firms that actually run models in response to user requests. When a user sends a prompt to an API, it is the inference provider's servers that process the input and return the model's output. We can categorize inference providers into three groups. The first group consists of the model creators themselves. As discussed above, most major model creators operate their own inference services and host their models directly

for customers. The second group comprises major cloud platforms, such as Amazon Web Services and Google Cloud. These platforms provide inference services for both open- and closed-source models, with closed-source models often subject to exclusive or preferential access arrangements. Little is known about the structure and terms of these contracts.

The third category is a growing set of specialized inference companies—firms like Together AI, Cerebras, and Groq— that focus on serving open-source models. These firms compete by optimizing hardware and software to deliver faster response times (lower latency), higher processing speeds (greater throughput), and lower prices. The result is that popular open-source models are often available from a dozen or more competing providers, each offering different combinations of price, speed, and reliability. This stands in sharp contrast to closed-source models, which are typically available from just one or a few providers.

Inference providers, therefore, compete along different dimensions depending on the type of model they serve. For providers serving closed-source models, the model itself is a key source of differentiation. By contrast, for providers serving open-source models, competition shifts to governance and the quality of inference, including latency, throughput, reliability, and price. For major cloud providers, existing customer relationships create an additional source of competitive advantage. Firms whose data, infrastructure, and workflows are already embedded within a particular cloud ecosystem may face switching costs, making integration, security compliance, and bundled service offerings important dimensions of competition.

Customers are primarily enterprises that purchase API access to incorporate AI into their operations or products.[2] Enterprise usage broadly falls into two categories. The first is internal use: a firm uses OpenRouter or directly goes to an inference provider for its own workflows, for example, to automate document review, generate code, or power internal search. The second is the *application layer*: firms build products in which the LLM is a core component embedded in a customer-facing service, such as a legal-tech tool that analyzes contracts or a coding assistant that offers code completion. These LLM-powered applications can offer a single model or a portfolio of models to their users; when multiple models are available, the end user often selects the model. Across both categories, model choice reflects the nature of the underlying task: cost-sensitive, high-volume workflows such as those used for e-commerce search gravitate toward cheaper and faster models, while high-stakes and more complex workflows such as vulnerability detection in software gravitate toward expensive frontier models.

The application layer is highly dynamic. A large ecosystem of startups and incumbent

[2]Individual users may also create accounts and access these APIs directly on most platforms. However, such usage is relatively uncommon, as most consumer-facing products operate on a subscription basis.

firms is developing AI-powered applications across a wide range of categories, including accounting, law, medicine, marketing, customer support, and education. The most mature segment to date is coding, where developer-assistant tools have achieved substantial adoption. The best-known third-party tools are GitHub Copilot and Cursor, which integrate access to many frontier models within a coding assistant interface. These compete with vertically integrated offerings such as Codex from OpenAI and Claude Code from Anthropic, which serve their own models. Applications in this layer differ not only in the models they provide access to, but also in how they are designed, how well they fit into a user's existing tools and processes, what features they offer to larger organizations (such as security controls and team administration), how they price their service, and the quality of their customer support. As a result, competition at the application layer involves both model performance and non-model product differentiation.

For most enterprises, entry into this ecosystem begins by signing up directly with an inference provider and accessing models through standardized API contracts with usage-based pricing. At this scale, relationships are typically governed by online terms of service rather than bespoke agreements. For high-volume users, however, procurement more closely resembles traditional enterprise IT contracting: large firms often negotiate individualized agreements with inference providers that specify volume discounts, committed capacity (a guaranteed amount of computing reserved for the buyer), service-level guarantees (commitments on uptime and response speed), data governance provisions, and security compliance requirements.[3]

While creating an account with any single inference provider is straightforward, managing accounts across multiple providers and customizing queries for each can be operationally costly for firms that wish to compare prices, optimize performance, or switch across models. To address this problem, a fourth layer has recently emerged: the *aggregator layer*, which sits between the inference layer and consumers and provides a single standardized API that enables access to models from multiple providers. Beyond consolidating access, aggregators can dynamically route requests to the lowest-cost or fastest provider of a given model and offer automatic fallback during outages, thereby reducing switching frictions and intensifying competition within the inference layer. The leading aggregator is OpenRouter, which offers access to hundreds of models from dozens of providers. Although this segment is expanding rapidly, most tokens are still processed directly by inference providers rather than through aggregators.

[3]In addition, a separate enterprise-services segment has emerged in which AI firms—or their implementation partners—work with clients to fine-tune models, build firm-specific AI agents, and integrate systems into proprietary workflows. In such cases, the enterprise may contract directly with the model provider or with a third-party integrator that builds and maintains customized applications on top of underlying API access.

Tokens, Pricing, and How Firms Pay

The fundamental unit of transaction in this market is the token. When a firm sends a request to an LLM through an API, it sends a sequence of prompt tokens (the input), and the model returns a sequence of completion tokens (the output). Providers charge per million tokens, with separate rates for prompt and completion tokens. Completion tokens are typically more expensive because generating output is more computationally intensive than processing input. This token-based pricing model means that AI inference is purchased much like a utility — firms pay for what they use, with no fixed costs for hardware or model deployment. Costs scale linearly with usage: a firm that doubles the number of documents it processes simply pays for twice as many tokens.

Inference providers post their prices, and most customers are price takers. Supply and demand are balanced primarily through non-price mechanisms. The primary channel is quality degradation: when demand exceeds capacity, latency rises and throughput falls, effectively rationing access by reducing service quality rather than raising prices. A second mechanism is batch processing, in which users submit requests and the provider processes them within 24 hours at a discounted rate, shifting demand to off-peak capacity. While prices can and do change over time, real-time price adjustments are not widely used to balance supply and demand.

Pricing varies enormously across models and providers. Frontier closed-source models—the most capable systems available—may charge $15 or more per million completion tokens, while commodity open-source models can cost less than a cent for the same volume. This variation reflects several factors: the model's intelligence, whether it is open- or closed-source, the provider's speed and reliability, and the competitive dynamics of the specific market segment.

In addition to prompt and completion tokens, many models feature "reasoning tokens"—tokens required for the model's internal deliberation steps—and "cache tokens," which allow users to reuse frequently sent prompts at a discount. These pricing primitives create a schedule in which firms can optimize their costs by choosing models, providers, and prompt strategies that match their particular usage patterns.

To make the pricing concrete, consider an example of a lawyer who submits a 50-page contract to Claude Opus 4.6 for a one-page summary. The text in the contract constitutes the prompt tokens, and the model's summary output constitutes the completion tokens. Assuming roughly 500 tokens per page, the input tokens cost about 13 cents (25,000 tokens at $5 per million) and the output tokens cost about one cent (500 tokens at $25 per million). Since Opus 4.6 is a reasoning model, the lawyer also pays for reasoning tokens, which cost

$25 per million tokens, though the number of reasoning tokens used is less predictable and varies by model. Now suppose the lawyer follows up with a question about liability exposure across multiple clauses. At this point, the original contract is already stored in the model's context cache[4]: she pays a discounted rate for the cached tokens (25,000 tokens at $0.50 per million), a standard prompt rate for the text of the follow-up question, and completion tokens for the response itself.

Data Sources

We study this market using data from OpenRouter, the leading LLM aggregator. OpenRouter provides a standardized API for accessing hundreds of models from dozens of providers, and its usage data allow us to observe token consumption by model, date, and by application category (e.g., programming, marketing, translation). We also observe pricing data for each model and provider, as well as information on model characteristics such as intelligence benchmarks, latency, and throughput. OpenRouter data are available from July 2023, with a balanced daily panel from January 2025 through December 2025.[5]

OpenRouter provides API access to virtually all commercially relevant LLMs and introduces new models immediately after creators release them. The prices observed in the OpenRouter data reflect the actual posted API prices charged by providers. As a result, the data provide an accurate view of the evolving supply side of the market—pricing, entry, and model turnover—without requiring us to reconstruct these features from multiple proprietary sources. On the demand side, we observe the aggregate usage of each model by provider at the daily level. Since OpenRouter accounts for only a small share of total API usage and skews toward developers and technology firms, we do not interpret our results on model demand as estimates of aggregate market shares, but instead use these data to study relative differences across models and use cases, and changes over time.

3 The Explosion of Supply

The Explosion of Models and Creators

Figure 2 summarizes the growth of the market along three dimensions: models, providers, and intelligence. Panel (a) shows the cumulative number of models available through OpenRouter over time, separated into open-source and closed-source categories. In early 2024, 63 models were available. By December 2025, that number increased to 668. The expansion is driven

[4] A context cache is a temporary memory that stores recently submitted text, allowing the model to reference it without reprocessing.

[5] We obtained this data via a scrape of the website, and are making it available as part of the replication package.

overwhelmingly by open-source models: as of December 2025, there are 449 open-source models compared to 219 closed-source models. Importantly, this growth reflects not just a few prolific creators releasing many models: the number of distinct model creators[6] has nearly doubled, from 47 at the start of 2025 to 88 by December, with the vast majority being open-source creators (78 open-source versus 10 closed-source). The number of closed-source creators has been stable at 10 since April 2025.

This proliferation of open-source models has important economic implications. In traditional technology markets, high fixed costs and proprietary standards tend to limit entry. The open-source model—in which creators bear the training costs but make weights freely available—dramatically lowers the barrier to downstream competition. Meta invested billions in training its Llama models but gave them away (Nagle and Yue, 2025). A Chinese model creator, DeepSeek, operating in a different regulatory environment, does the same. Any firm with GPUs (graphics processing units—specialized computer chips used for AI computation) can begin serving an open-source model, and anyone with machine-learning expertise can fine-tune or distill an existing model for a specialized application.

A key question is whether it is worthwhile for firms to create open-source models, given the presence of external inference providers that compete to monetize them. Several reasons have been given for open-sourcing, including that open-source models are complements to algorithmic recommendations (e.g., for Meta) and hardware (e.g., for Nvidia), and that open-source models hurt competitors with closed-source models. These open-source releases function as a competitive subsidy to the rest of the market, keeping prices and entry barriers low in a way that has little parallel in other capital-intensive technology industries. The result is an ecosystem in which innovation occurs not only at the frontier but also in the adaptation and deployment of existing capabilities.

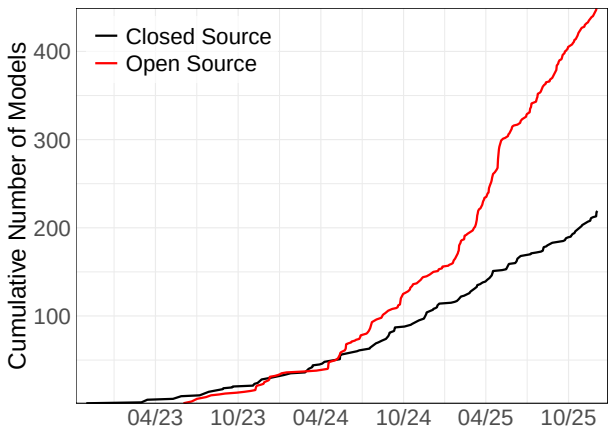
The Rise of Inference Providers

Panel (b) documents the rapid entry of inference providers. In November 2024, there were 27 providers on OpenRouter. By December 2025, that number had grown to 90. This entry is concentrated in the provision of open-source models: popular open-source models are often hosted by a dozen or more competing providers. In contrast, closed-source model creators often serve their models from just a few providers, with new providers typically gaining access through investment partnerships with the creator.

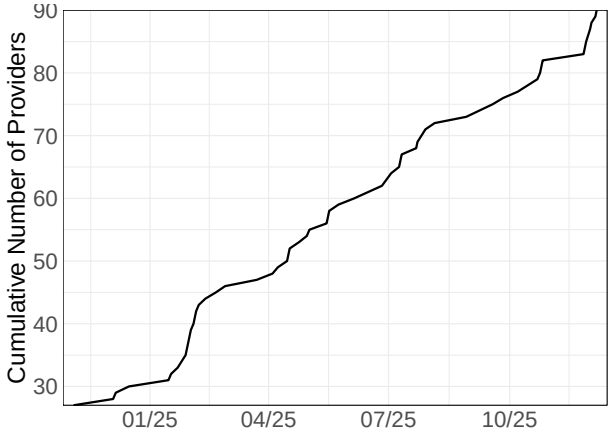
When multiple providers offer the same underlying model, they can differentiate on other

[6] A "creator" is the organization that publishes the model, identified by OpenRouter's publisher field (e.g., Anthropic, Google, Meta, DeepSeek). This corresponds to the corporate brand releasing the model rather than a legal taxpayer entity, and generally to the corporate parent.

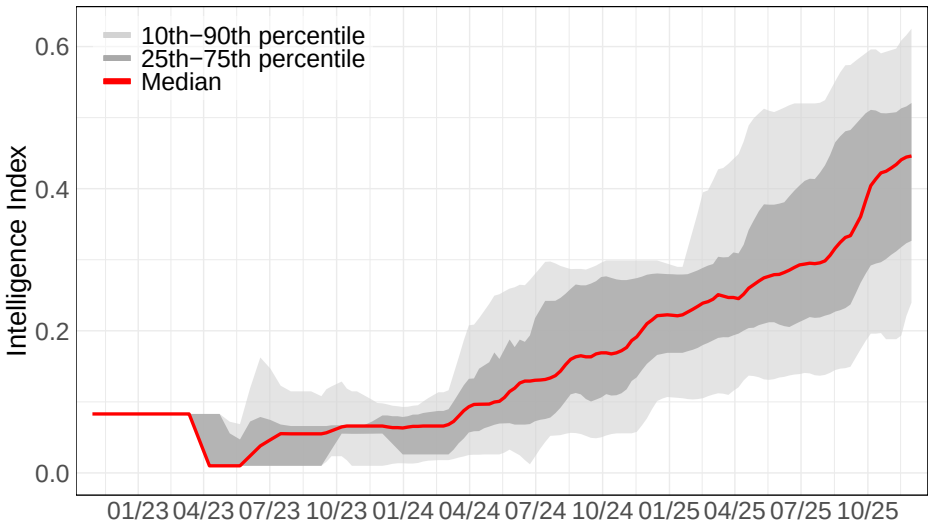
Figure 2: Rapid Growth in Models, Providers, and Intelligence



(a) Number of Models, by Type



(b) Number of Inference Providers



(c) Intelligence of Newly Released Models Over Time

Notes: Panel (a) shows the cumulative number of models available on OpenRouter, separated by open- and closed-source. Panel (b) shows the cumulative number of inference providers. Panel (c) shows the distribution of the Intelligence Index for models released within six months of each date. The red line is the median; shaded areas show the interquartile and 10th–90th percentile ranges. Source: OpenRouter and Artificial Analysis.

dimensions: price, latency, and reliability. Our data show substantial variation across these dimensions. For example, among providers hosting the same open-source model, throughput can vary by a factor of five, and prices can differ by more than an order of magnitude (Demirer et al., 2025). These seemingly homogeneous products (the same model weights) are sold by differentiated sellers, with price and non-price competition. The contrast with closed-source models is stark: their prices remain largely rigid over their lifetimes and there

is almost no variation in price across providers as model creators can dictate prices through contractual arrangements (Demirer et al., 2025).

Measuring Capabilities

A central challenge in the market for LLM inference is measuring differences in model quality. To address this challenge, the industry has developed *benchmarks*—standardized evaluation suites that test models on tasks such as mathematical reasoning, coding, and scientific problem solving and assign quantitative performance scores. These benchmarks are typically highly correlated across domains, suggesting a dominant general capability factor rather than isolated task-specific strengths (Papailiopoulos, 2026). This general capability may stem from model or training data size, consistent with well-documented scaling laws (Kaplan et al., 2020).

While benchmarks are useful, they are imperfect since real-world enterprise tasks may differ from exam-style evaluations. In what follows, when we discuss model capabilities, we measure them using the composite Intelligence Index reported by Artificial Analysis.[7] It is important to note that other model characteristics such as latency, modality (image, audio, text), and cost may determine which model is used for a particular task, especially since some tasks do not require frontier intelligence.

Improving Intelligence

Panel (c) of Figure 2 tracks the capabilities of newly released models over time, where "new" is defined as released within the preceding six months. The red line plots the median intelligence of new releases, while the shaded regions depict the 25th–75th and 10th–90th percentile ranges. The median intelligence of new models has risen from about 0.1 when GPT-3.5 launched to above 0.4 today—roughly a fourfold improvement. This represents a shift from models that frequently hallucinate and struggle with multi-step reasoning to models that can solve complex problems and have much lower rates of hallucination (Kalai et al., 2025). The distribution of intelligence across newly released models has also widened: the best models today score roughly six times higher than the earliest models in our sample, yet firms continue to release models with lower capabilities as well. This improvement in capabilities is a supply-side story with demand-side consequences. As models become more capable, they become useful for a wider range of tasks, potentially expanding the total addressable market for AI services. At the same time, the rapid pace of improvement creates

[7] Artificial Analysis is an independent evaluation platform that benchmarks AI models across multiple dimensions. Its Intelligence Index combines ten standardized evaluations spanning reasoning, knowledge, mathematics, and coding.

strong incentives for firms to stay near the frontier.

Taken together, the evidence indicates that the industry is in an explosive expansion phase. Entry occurs simultaneously across multiple layers of the supply chain—model creation, inference provision, and application development—while rapid improvements in capabilities continually reshape the competitive frontier. Open-source releases play a central enabling role, lowering downstream entry barriers and accelerating experimentation. The result is a layered and quickly changing ecosystem in which specialization, vertical integration, and horizontal differentiation coexist, and in which competitive conditions can shift quickly as new models and providers enter.

4 The Price of Intelligence

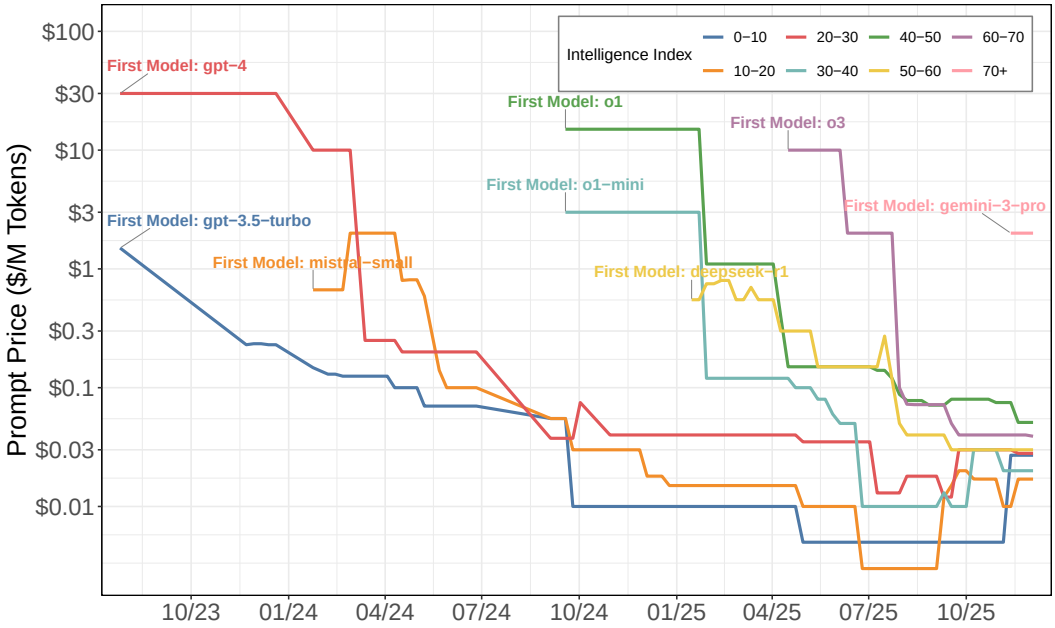
Perhaps the single most striking fact about the business-to-business AI market is the speed at which the price of intelligence has fallen. To be sure, plunging prices in some form have been seen with other technologies. The deregulation of telecommunications in the 1990s cut the price of a long-distance phone call by about 90% over 15 years. Gordon Moore, who would later be a co-founder of Intel, predicted back in 1965 when working for Fairchild Semiconductor that the number of transistors on a chip would double every two years (Moore, 1965). Moore's Law delivered roughly a hundredfold reduction in the cost of computing per decade. The LLM market has compressed an even larger price decline, which might fairly be characterized as a price collapse, into just two years.

The 1,000-Fold Price Decline

Figure 3 shows the minimum price for models at each level of intelligence (the vertical price axis is a log scale). Each line represents a distinct capability tier, measured by the Intelligence Index. In mid-2023, GPT-4-class intelligence cost roughly $30 per million prompt tokens—the only model at that capability level. By late 2025, the same level of intelligence is available for about three cents, a decline of approximately 1,000-fold over two years. The pattern is not limited to GPT-4-class models: steep price declines are visible across all intelligence tiers.

Three forces drive these declines. First, advances in model architecture and training techniques allow creators to achieve the same level of intelligence with fewer neural network parameters. This translates into less computation required per query. Second, improvements in hardware efficiency and serving infrastructure enable inference providers to deliver a given model at lower marginal cost, reducing the price of inference even while holding model capability constant. Finally, competition—both between closed-source providers and from new open-source models—puts downward pressure on prices.

Figure 3: The Falling Price of Intelligence



Notes: This figure shows the minimum prompt price per million tokens 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. Source: OpenRouter and Artificial Analysis.

The implications of this price decline for firm adoption are profound. Tasks that were prohibitively expensive to automate a year ago are now economically viable, and the set of applications for which AI is cost-effective is expanding rapidly. At the same time, the average price that firms actually pay per token has remained relatively stable over this period (Demirer et al., 2025). Firms are not simply buying the same intelligence more cheaply—they are upgrading to more capable models as prices fall. The intelligence of models in use has risen steadily, with the token-weighted median intelligence score increasing from 0.30 in January 2025 to 0.44 by year-end (Demirer et al., 2025). This pattern is consistent with firms using price declines to purchase better, not just cheaper, AI. The dynamic echoes the history of personal computers and smartphones—modest price declines paired with order-of-magnitude capability gains—except that for LLMs, both prices and capabilities have moved sharply in the buyer's favor.

The Price–Intelligence Relationship

Figure 4 shows the relationship between price and intelligence across all models as of December 2025. Each point is a model; color indicates whether the model is open-source (orange) or closed-source (blue); and point size reflects total token usage.

Two patterns stand out. First, the relationship between price and intelligence is approximately linear in logs, meaning that it is highly nonlinear in levels. Moving from a mid-tier

to a frontier model is associated with a disproportionately large increase in cost. This is consistent with the economic intuition that producing the last increment of intelligence is far more expensive than producing the first. Second, there is enormous price dispersion at any given level of intelligence. For models with similar benchmark scores, prices can vary by more than two orders of magnitude.

Open-Source vs. Closed-Source Models

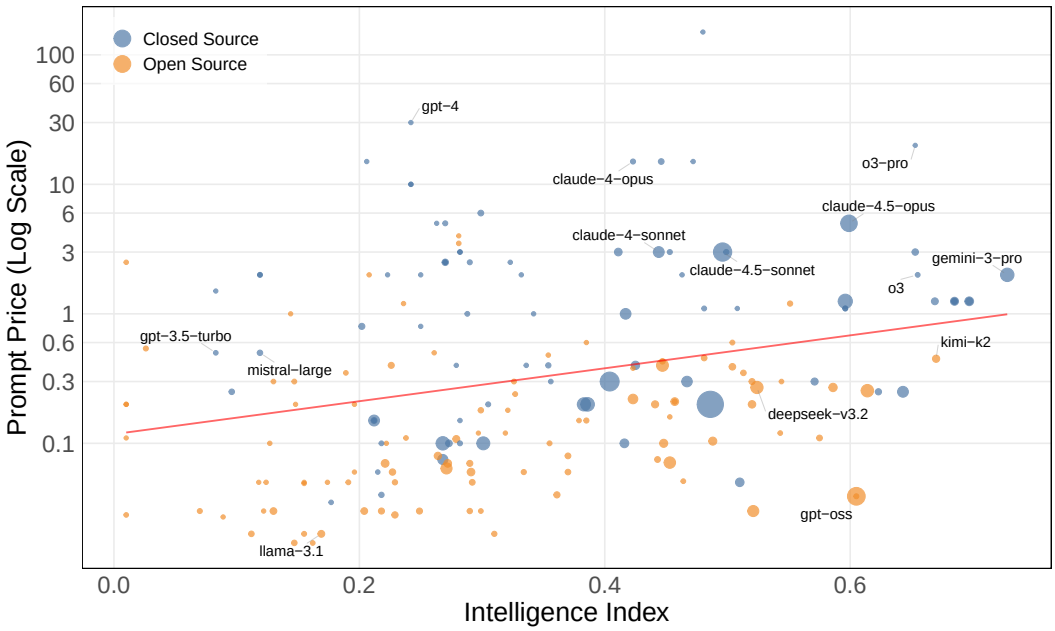
Much of the price dispersion turns on a fundamental distinction in how models are distributed. *Closed-source* models, such as OpenAI's GPT series and Anthropic's Claude, are proprietary: the creator keeps the underlying model weights secret and sells access only through its own API or a small number of authorized partners. Because the creator controls the sole channel to the model, it can set the price directly, and competition for that specific model is limited. *Open-source* models, such as Meta's Llama and DeepSeek's releases, are distributed differently: the creator publishes the model weights, allowing any firm with sufficient computing capacity to host and serve the model. This means a single open-source model is typically offered by many competing inference providers, and price competition among them drives the cost of serving that model down toward marginal cost.

Controlling for intelligence level, open-source models are approximately 90% cheaper than their closed-source counterparts. Despite the price advantages, the open-source models are used far less than their closed-source counterparts, as indicated by the point sizes in Figure 4, which scale with token volume. Overall, open-source models account for less than 30% of total token consumption.

Why would firms pay substantially more for closed-source models when open-source alternatives with comparable benchmark performance are available at a fraction of the cost? Several explanations are plausible, and they are not mutually exclusive.

First, benchmarks may not fully capture the dimensions of quality that firms care about. Closed-source creators invest heavily in "reinforcement learning from human feedback" (often abbreviated as RLHF), a training process in which human evaluators rank model outputs to iteratively improve response quality. They also invest in "safety alignment," which shapes how a model reflects human values and ethics. These post-training refinements may improve the model's usefulness for real-world business tasks in ways that benchmarks may not measure. Second, closed-source models come with contractual guarantees, enterprise support, and service-level agreements that reduce the risk to firms integrating AI into critical business processes. Third, adjustment costs may play a role to the extent that a firm that has optimized its prompts, workflows, and internal tooling around a particular model faces costs in

Figure 4: Price vs. Intelligence, Open-Source vs. Closed-Source



Notes: Scatter plot of prompt price (log scale) against the Intelligence Index as of December 6, 2025. Point size is proportional to total token usage. Color indicates open-source (orange) vs. closed-source (blue) models. The fitted line shows the average price–intelligence relationship. Source: OpenRouter and Artificial Analysis.

migrating to an alternative, even if the alternative is cheaper. Fourth, firms may be to some extent mistaken in perceiving open-source models as inferior substitutes for closed-source alternatives, even when actual performance is similar—a belief that, whether justified or not, can sustain the price premium.

The coexistence of expensive closed-source and cheap open-source models is one of the most economically interesting features of this market. It suggests that the market for AI services is not a simple commodity market in which the lowest-cost producer wins. Instead, it is a differentiated market in which firms trade off multiple attributes—intelligence, price, reliability, brand trust, and integration costs—when choosing a model.

5 Market Dynamics and Differentiation

So far, the market resembles a race in which some creators compete to advance the frontier of intelligence and charge a premium, while other creators offer more targeted models optimized for specific tasks. Over time, one possibility given the number of models and falling prices is that this market results in a race to the bottom, with intelligence becoming a commodity priced close to marginal cost. However, this outcome has not happened yet.

An important aspect of this market is heterogeneous demand across use cases. A software firm generating code, a law firm summarizing depositions, and a retailer translating product

listings face different task requirements, latency constraints, reliability needs, and willingness to pay. As a result, technological progress does not collapse the market into a single dominant product.

A Market in Motion

To illustrate these dynamics, we divide the market into a dozen use cases: trivia, translation, technology, science, roleplay, programming, search engine optimization (SEO), marketing, legal, health, finance, and academia. Figure 5 makes these dynamics visible. Each row corresponds to a use-case category, and the horizontal axis spans the weeks in 2025. At each point in time, the color indicates which creator's model holds the largest share of usage in that category, while the labels within each segment identify the specific model leading during that period. We highlight the top six creators—Google, Anthropic, OpenAI, DeepSeek, xAI, and Meta—and group all others into an "Other" label.

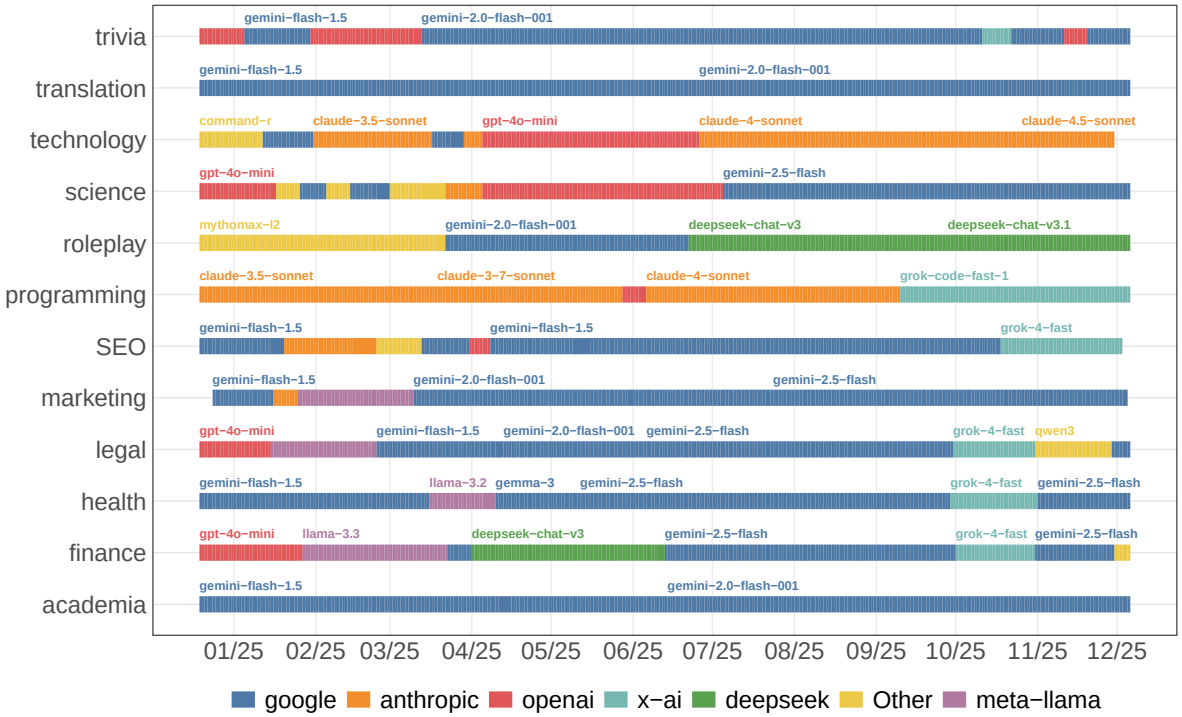
Two patterns stand out. First, no single creator dominates across all categories: the figure is a mosaic of colors rather than a monochrome. Second, the dominant color within any row shifts frequently as new model releases displace incumbents. The leading model in the legal category changed multiple times over the year; translation, marketing, and finance followed similarly volatile patterns.

The churn is even more pronounced at the model level. Over this period, several major model launches occurred, and across categories, leading models were frequently replaced by newly released versions. In programming, for example, the leading model moved quickly from Claude Sonnet 3.5 to 3.7, and then to 4 and 4.5, closely following their release dates. Similarly, in categories where Google leads, we observe transitions toward newer frontier models. Importantly, adoption is not uniform across use cases. Some categories—such as programming and legal—transition rapidly to newly released models, whereas others—such as trivia and marketing—continue to rely on earlier versions.

Horizontal Differentiation

The cross-category patterns of which model leads also reveal horizontal differentiation in addition to the vertical innovation described above. Different use cases require different combinations of capability, speed, reliability, and price. A developer using a coding assistant needs a model that can reason through complex logic and generate correct code—errors are immediately apparent when the program fails to run. A marketing team generating ad copy places greater weight on tone, creativity, and stylistic control. A customer-service chatbot must prioritize speed and low cost, as it processes large volumes of relatively low-value queries. These differing requirements steer buyers toward different regions of the

Figure 5: Market Leadership Shifts Across Use Cases



Notes: Each row is a use-case category. The underlying data are weekly: each segment within a row corresponds to a week in 2025, while the horizontal axis is labeled by month. Color indicates the creator of the model holding the largest share of prompt tokens in that category during that week (30-day rolling window). Frequent color changes within rows reflect the leading model changing rapidly following new releases. Source: OpenRouter.

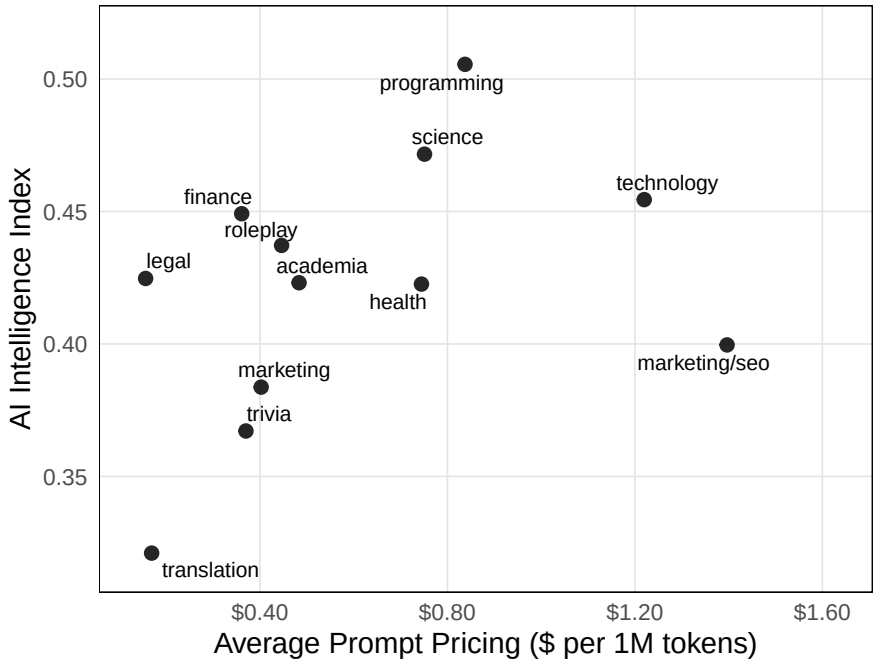
capability–price frontier, creating persistent segmentation.[8]

Figure 6 documents this phenomenon by plotting the average AI Intelligence Index (y-axis) and prompt pricing (x-axis) by category. Programming is the most intelligence-intensive application, with an average score of 0.51, followed by science at 0.47. At the other end, translation (0.32) and trivia (0.37) rely on models with substantially lower intelligence scores. The variation in price paid across categories is substantial: SEO applications pay an average of $1.40 per million prompt tokens, while legal and translation applications pay just $0.16 and $0.17, respectively—a nearly ninefold difference. Applications that demand higher capability tend to pay more, while high-volume or simpler tasks gravitate toward lower-priced models.

This interaction between vertical and horizontal differentiation helps explain several features of the market. It explains why many creators can coexist: a firm that produces a lower-intelligence model at a low price is not simply losing to more capable competitors—it

[8]These patterns are also visible beyond OpenRouter. For example, Anthropic's Claude models are known to be especially popular among software developers: coding tasks account for the largest share of usage on Anthropic's own platform (Handa et al., 2025), and enterprise surveys rank Claude models first in code generation (Menlo Ventures, 2025).

Figure 6: Intelligence and Pricing Vary Across Use Cases



Notes: The figure shows the usage-weighted average intelligence score and prompt price per million tokens used in each category over a 30-day period (November 7 to December 6, 2025). Source: OpenRouter and Artificial Analysis.

is serving a different segment of demand. It explains why frontier models, despite their impressive capabilities, account for only a modest share of total token consumption (as illustrated earlier in Figure 4): for many applications, the incremental intelligence does not (yet) justify the higher cost. And it explains why the market has not converged on a single winning model, despite the fact that one model is measurably "best" on standardized benchmarks at any given moment.

These results have two important implications. First, model releases function as discrete competitive shocks that can rapidly reallocate demand. Because code-driven switching costs are minimal for most applications—an API call can be redirected without new infrastructure investment—users adjust quickly when relative performance shifts. Whether users do switch also depends on whether their workflows are generic rather than being optimized for only one model. Second, they imply that any competitive advantage derived from having the best model at a given moment is inherently temporary. The returns to innovation are fleeting, creating strong incentives for continuous investment in model improvement.

The dynamism we document is not without precedent. Early technology markets routinely exhibit turbulent leadership before consolidation sets in. In the early American automobile industry, hundreds of manufacturers competed in the first two decades of the twen-

tieth century, with market leadership shifting frequently as engine designs, body styles, and production methods evolved rapidly; by the 1930s, the industry had consolidated around a handful of dominant firms (Klepper, 1996). The early search engine market followed a similar pattern: AltaVista, Lycos, Excite, and Yahoo all held leading positions in the late 1990s before a superior entrant displaced them within a few years. In both cases, the early phase was characterized by rapid entry, frequent leadership changes, and users' quick adoption of better alternatives — precisely the pattern we observe in LLMs today.

What is distinctive about the LLM market is the pace: the turnover we document occurs over months rather than years, reflecting both the speed of model development and the small cost of switching models via standardized APIs. The market's segmentation across use cases, intelligence levels, and price points also suggests that, if consolidation occurs, it is unlikely to yield a single dominant model in the enterprise LLM market. Different buyers have distinct needs, and different creators have identified niches to serve them. Whether the LLM market will eventually consolidate into a small number of dominant platforms, as the automobile and search markets did, or sustain a more fragmented equilibrium, remains an open question.

6 AI Adoption and Cost in Historical Perspective

The economic literature on general-purpose technologies documents a recurring pattern: widespread adoption and productivity effects arrive slowly, often decades after invention (Bresnahan and Trajtenberg, 1995). Steam power and electrification took 40 years or more to diffuse broadly across the economy; computers and information technology took roughly 25 years from early commercial availability in the 1970s to pervasive business adoption in the mid-1990s.

To understand why these lags occur and whether the trajectory of AI technology might be different, it is useful to decompose adoption costs into two categories. The first is the *cost of access and usage*—what firms must pay before and while using the technology. Electrification required connection to a power grid that took decades to build, and electricity itself remained expensive for decades (David, 1990). The internet required networking infrastructure, hardware, and technical expertise. Even after firms made the necessary investments to gain access to these new technologies, high usage costs limited the range of applications that were economically viable. Technologies diffuse widely only after both barriers fall: the technology must be accessible, and usage must be affordable.

The second category is *organizational adjustment costs*—the softer, often slower process by which firms and workers learn to use a new technology effectively. Even after electricity was widely available and affordable, factories took a generation to reorganize their floor

layouts around distributed electric motors, rather than central steam shafts (David, 1990). The productivity gains from information technology similarly lagged adoption by a decade or more, as firms redesigned workflows, retrained workers, and developed new managerial practices (Brynjolfsson and Hitt, 2000). These organizational co-invention costs are more difficult to observe and often slower to resolve, as they depend on firm-level capabilities, managerial practices, and industry-specific constraints that vary across sectors and use cases.

Artificial intelligence, delivered through APIs, faces a fundamentally different cost structure. The physical infrastructure required for access—the internet, cloud computing platforms, electricity grids, and connected devices—already exists. A firm that wants to adopt LLMs does not need to build roads or string wires; it needs an API key and a few lines of code. This means that *access* to AI can diffuse far faster than access to electricity or the internet.

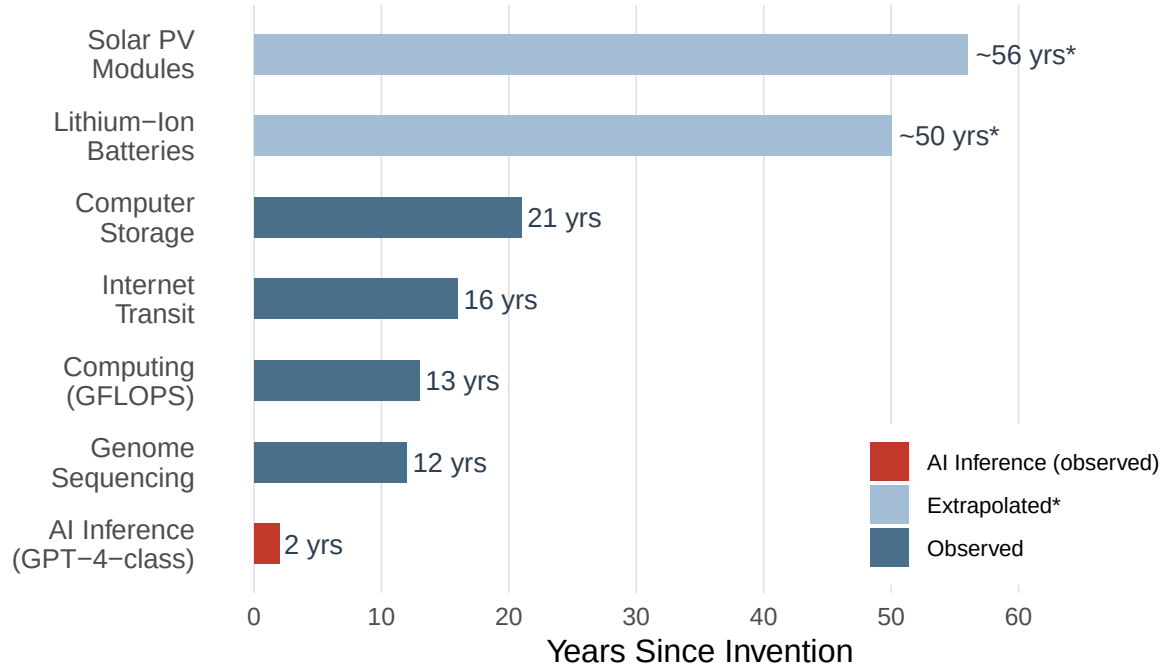
The cost of usage has also collapsed at an unprecedented rate. Figure 7 compares how quickly the prices of several foundational technologies declined, measured as the number of years required for a 1,000-fold fall in price. AI inference stands out as an extreme case: as noted earlier, the price of GPT-4-class inference fell roughly 1,000-fold in about two years, far faster than internet transit, genome sequencing, computing, or computer storage, and vastly faster than the trajectories implied by batteries and solar. The broader pattern is that newer digital technologies can experience extraordinarily rapid cost compression. For AI, this means that both components of the cost barrier—the cost of access and the cost of usage—have fallen faster than for any comparable technology.

The low costs have encouraged rapid early adoption. A survey conducted in early 2026 reports that 43% of US workers have used AI in their jobs (Bick et al., 2026). And if we look at country-level adoption, which is commonly used to study technology diffusion (as in Comin and Hobijn, 2010), LLMs have already diffused: our data indicate usage in more than 200 countries and territories. No previous general-purpose technology has achieved this breadth of access so quickly.

However, the category of organizational adjustment costs remains, and early evidence suggests it may now be the binding constraint on realizing the economic potential of AI technologies. The same survey that documents widespread worker-level adoption finds that productive integration—using AI in the actual production of goods and services rather than for experimentation or personal assistance—lags substantially. Only 7% of US firms and 4% of European firms report using AI in production (Bick et al., 2026).[9] The gap between

[9]"Production" here refers specifically to the use of AI in producing goods or services, as defined by the US Census Bureau's Business Trends and Outlook Survey, which covers nonfarm employer businesses across all sectors and firm sizes. It excludes AI use in other business functions such as marketing, logistics, or

Figure 7: Time to 1,000-Fold Fall in Price



Notes: Each bar shows the number of years required for a technology's price to fall by a factor of 1,000. Solid bars denote technologies for which a 1,000-fold decline has been observed; light bars marked with * show extrapolations at the observed constant rate of decline for technologies that have not yet reached the 1,000-fold threshold. Sources: OpenRouter data (Aubakirova and Midha, 2025); NHGRI genome sequencing costs via Our World in Data (National Human Genome Research Institute, 2022; Dattani et al., 2023); cost per GFLOPS (billions of floating-point operations per second, a measure of computing speed) (AI Impacts, 2017); internet transit prices (DrPeering, 2016); computer storage prices (Ritchie and Roser, 2023); solar PV module prices (Ritchie et al., 2024); lithium-ion battery prices (Ritchie and Roser, 2024).

diffuse worker-level adoption and narrow firm-level production use suggests that the binding constraint on AI's economic impact has shifted from access and cost to organizational transformation.

The level of organizational investment in adapting to the possibilities of AI technology is sharply heterogeneous across industries. In software development, where work is already conducted through code in standardized environments—the same interfaces that LLMs use—integration costs are low, and productivity gains have already materialized: studies of AI-assisted coding find that developers complete tasks substantially faster and produce more code output (Peng et al., 2023; Sarkar, 2025). In other settings—healthcare,

administration. This narrow definition accounts for much of the apparent disconnect between worker-level and firm-level adoption rates: when the Census broadened the question in November 2025 to cover any business function, the share of US firms using AI rose to 17%, and Bick et al. (2026) estimate that on a basis comparable to European firm surveys it would be roughly 34%. Two further factors contribute to the remaining gap. The firm-level rate is an unweighted share of firms and thus dominated by small businesses, whereas AI-using workers are concentrated in large firms. In addition, some workers use AI at work without formal adoption by their employer.

law, finance, manufacturing—the organizational bottleneck is considerably more binding. A hospital system considering the possibility of adopting AI diagnostic tools, for example, must also redesign clinical workflows, obtain regulatory clearance, retrain staff, and address liability structures. A law firm would have to develop quality-control procedures and navigate evolving professional-responsibility guidance.

Moreover, the productivity gains from general-purpose technologies have accelerated once complementary products emerged that made reorganization practical—standardized electric motors and modular machine tools for electrification, enterprise software and networking equipment for information technology. In the artificial intelligence market, an analogous ecosystem is developing at a much faster pace. Increasingly, intelligence is bundled into software systems that perform complex, multi-step tasks on behalf of users. For example, tools such as Claude Code and Codex bundle model inference with execution logic and workflow integration, allowing users to delegate complex tasks rather than issue individual prompts. By reducing integration costs, these innovations expand the range of economically viable uses.

While the market described in this paper has developed the models and inference technologies and made them widely accessible, the complementary applications that translate these capabilities into usable products remain in their infancy, except for a few industries. Currently, a large ecosystem of startups—backed by billions of dollars in venture capital—is rapidly developing complementary applications. How quickly these markets produce compelling products—and how rapidly firms adapt their internal processes to integrate them—will be central to determining AI's ultimate economic impact.

7 Where Is the Market Headed?

The market for machine intelligence is only a few years old, yet it has already developed a layered supply chain, attracted dozens of competing model creators and nearly a hundred inference providers, and driven the quality-adjusted price of intelligence down by three orders of magnitude. How both the technology and the market evolve from here is a multi-trillion-dollar question and one that calls for humility.

A central question is whether and how soon the AI market may consolidate. Early signs of stabilization have already appeared: the set of frontier model creators has remained largely unchanged in the latter part of 2025 and into early 2026, even as innovation continues. If training costs keep rising, fewer firms will be able to compete at the frontier, given that state-of-the-art runs now cost hundreds of millions of dollars (Epoch AI, 2025). But these barriers to entry may not last. Open-source models are close behind, and improvements in algorithms may reduce training and deployment costs. There is also a possibility that local

devices, such as phones, will be able to serve at least some inference needs without relying on cloud providers. In any of these cases, costs may collapse even further.

For most of economic history, intelligence was bundled with human labor and allocated through labor markets. Today, machine intelligence is being unbundled from humans, supplied through its own markets, and purchased on demand. Most informed observers expect frontier capabilities to keep improving while prices continue to fall (Murphy et al., 2025). As intelligence becomes cheaper and more capable, it will become a standard input into production, and more tasks will be automated. The declining cost of intelligence may therefore alter the boundaries of firms and the structure of labor demand. It remains to be seen whether these changes quickly translate to productivity or whether bottlenecks slow this process down.

References

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

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

AI Impacts (2017). Trends in the Cost of Computing. Historical cost per GFLOPS, inflation-adjusted to 2013 USD. See also Wikipedia history of GFLOPS costs at https://aiimpacts.org/wikipedia-history-of-gflops-costs/.

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

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

Bick, A., A. Blandin, D. J. Deming, N. Fuchs-Schündeln, and J. Jessen (2026). Mind the Gap: AI Adoption in Europe and the US. Prepared for the Brookings Papers on Economic Activity.

Bresnahan, T. F. and M. Trajtenberg (1995). General Purpose Technologies: "Engines of Growth"? *Journal of Econometrics 65*(1), 83–108.

Brynjolfsson, E. and L. M. Hitt (2000). Beyond Computation: Information Technology, Organizational Transformation and Business Performance. *Journal of Economic Perspectives 14*(4), 23–48.

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

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).

Comin, D. and B. Hobijn (2010). An Exploration of Technology Diffusion. *American Economic Review 100*(5), 2031–2059.

Dattani, S., F. Spooner, H. Ritchie, and M. Roser (2023). Cost of Sequencing a Full Human Genome. Based on NHGRI data.

David, P. A. (1990). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. *American Economic Review 80*(2), 355–361.

Demirer, M., A. Fradkin, N. Tadelis, and S. Peng (2025). The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs. *NBER Working Paper* (34608).

DrPeering (2016). Internet Transit Prices—Historical and Projected. Wholesale IP transit pricing per Mbps, 1998–2015.

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

Epoch AI (2025). Parameter, Compute and Data Trends in Machine Learning.

Greenstein, S. (2015). *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton University Press.

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).

Kalai, A. T., O. Nachum, S. S. Vempala, and E. Zhang (2025). Why Language Models Hallucinate. *arXiv preprint arXiv:2509.04664* .

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* .

Klepper, S. (1996). Entry, Exit, Growth, and Innovation over the Product Life Cycle. *American Economic Review 86*(3), 562–583.

Klepper, S. (2002). Firm Survival and the Evolution of Oligopoly. *The RAND Journal of Economics 33*(1), 37–61.

Lerner, J. and J. Tirole (2005). The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives 19*(2), 99–120.

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

Moore, G. E. (1965, April 19). Cramming More Components onto Integrated Circuits. *Electronics Magazine 38*(8), 114–117.

Murphy, C., J. Rosenberg, J. Canedy, Z. Jacobs, N. Flechner, R. Britt, A. Pan, C. Rogers-Smith, D. Mayland, C. Buffington, S. Kučinskas, A. Coston, H. Kerner, E. Pierson, R. Rabbany, M. Salganik, R. Seamans, Y. Su, F. Tramèr, T. Hashimoto, A. Narayanan, P. Tetlock, and E. Karger (2025, nov). The Longitudinal Expert AI Panel: Understanding Expert Views on AI Capabilities, Adoption, and Impact. Report, Forecasting Research Institute.

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

National Human Genome Research Institute (2022). The Cost of Sequencing a Human Genome. NHGRI Genome Sequencing Program. Series covers 2001–2022.

Papailiopoulos, D. (2026). You Don't Need to Run Every Eval.

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).

Ritchie, H., P. Rosado, and M. Roser (2024). Solar PV Module Prices. Our World in Data. U.S. Module Prices from DOE/NREL.

Ritchie, H. and M. Roser (2023). The Price of Computer Storage Has Fallen Exponentially Since the 1950s. Our World in Data.

Ritchie, H. and M. Roser (2024). Battery Price Decline. Our World in Data.

Sarkar, S. K. (2025). AI Agents, Productivity, and Higher-Order Thinking: Early Evidence from Software Development. *Available at SSRN 5713646*.