Topics in the Economics of AI + The Market for Intelligence

Andrey Fradkin (Boston University / Amazon)

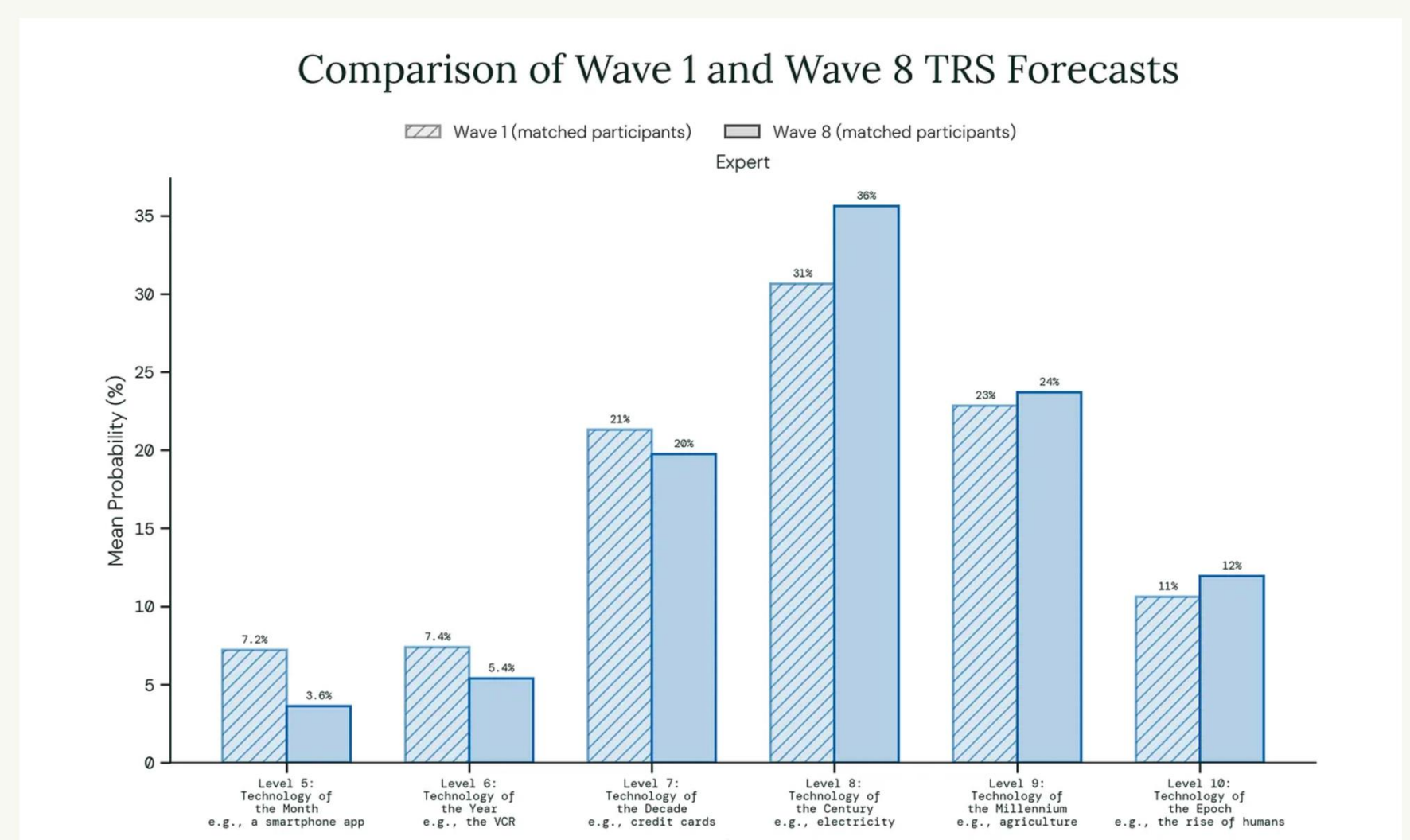
The views presented are solely the authors' and do not necessarily reflect the views of Amazon.

Some motivating facts that all economists should know about AI.

*Not based on my own research.

**Apologies if you know this already.

Increasingly many experts believe that AI will be the technology of the century / millennium / epoch.

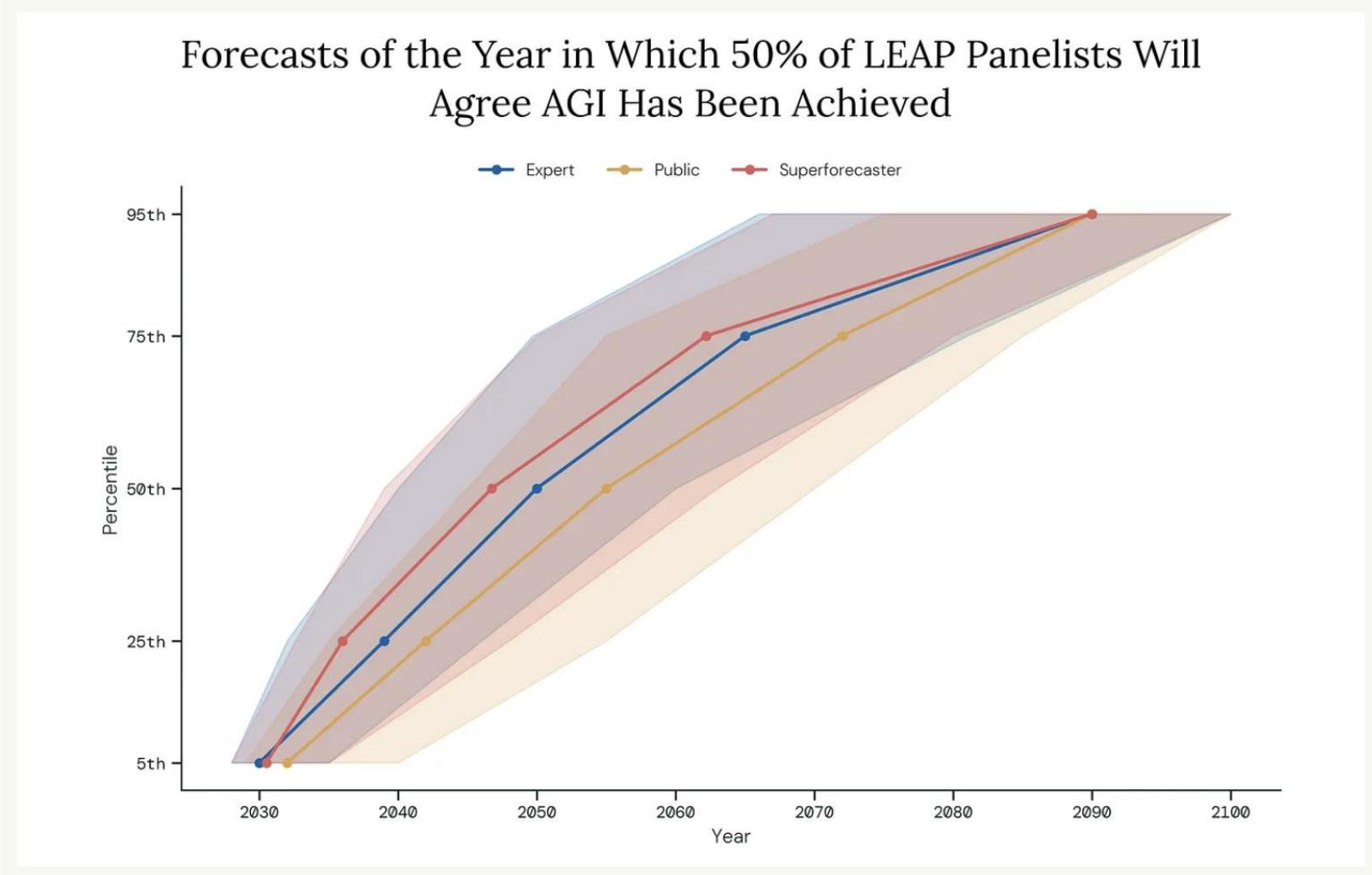


Source: Murphy et al. (2025). The Longitudinal Expert AI Panel.

Median estimate that Artificial General Intelligence (AGI) will be reached by 2050.

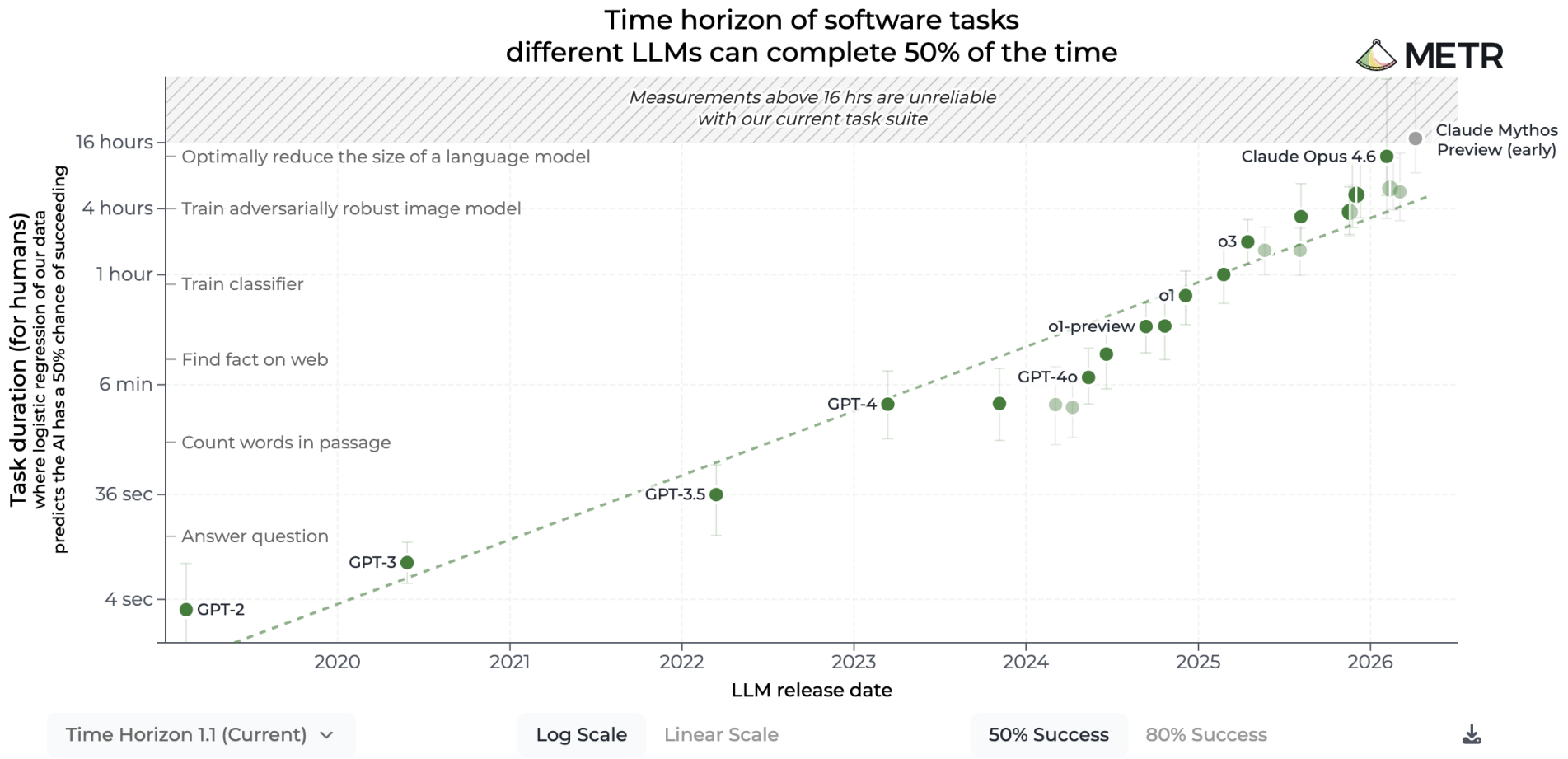
AGI Definition Used:

- Can outperform the 90th percentile professional human employee in every primarily non-physical occupation, across all sectors, on at least 90% of the economically useful non-physical tasks that they perform.
- Has an inference cost no more than 5x the cost of equivalent human labor.



Source: Murphy et al. (2025). The Longitudinal Expert AI Panel.

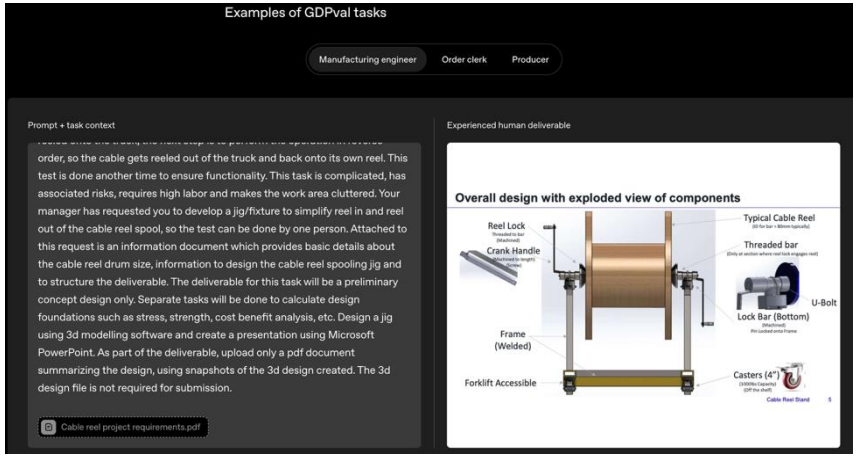
The Famous METR Plot – Time Horizon of Tasks

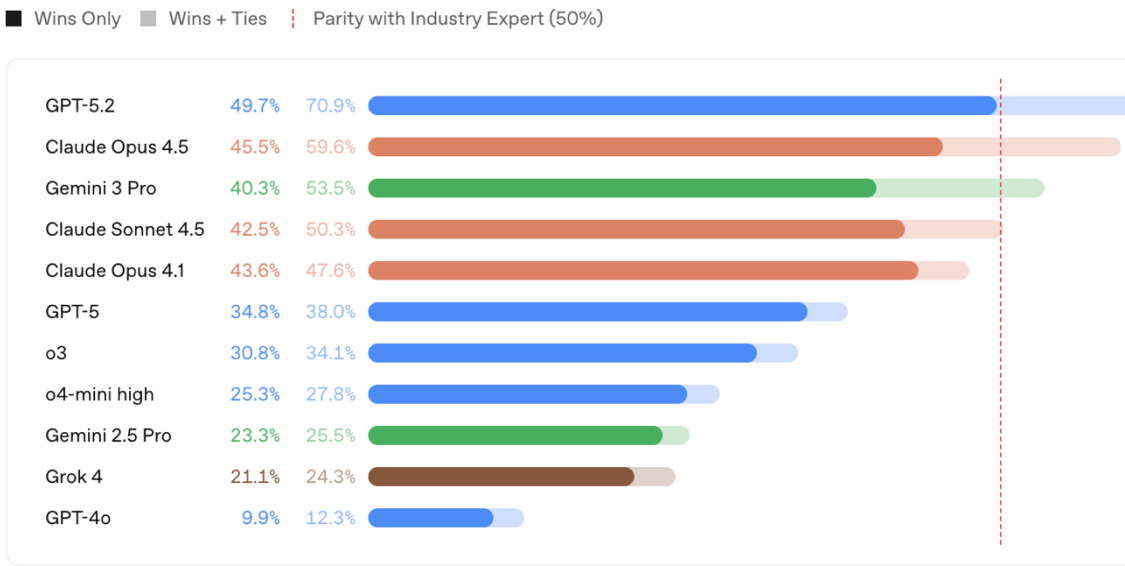


Source: METR. Measuring AI Ability to Complete Long Tasks.

GDPVal from OpenAI

- Occupations selected from the top 9 industries contributing to U.S. GDP.
- Humans rate the better output.

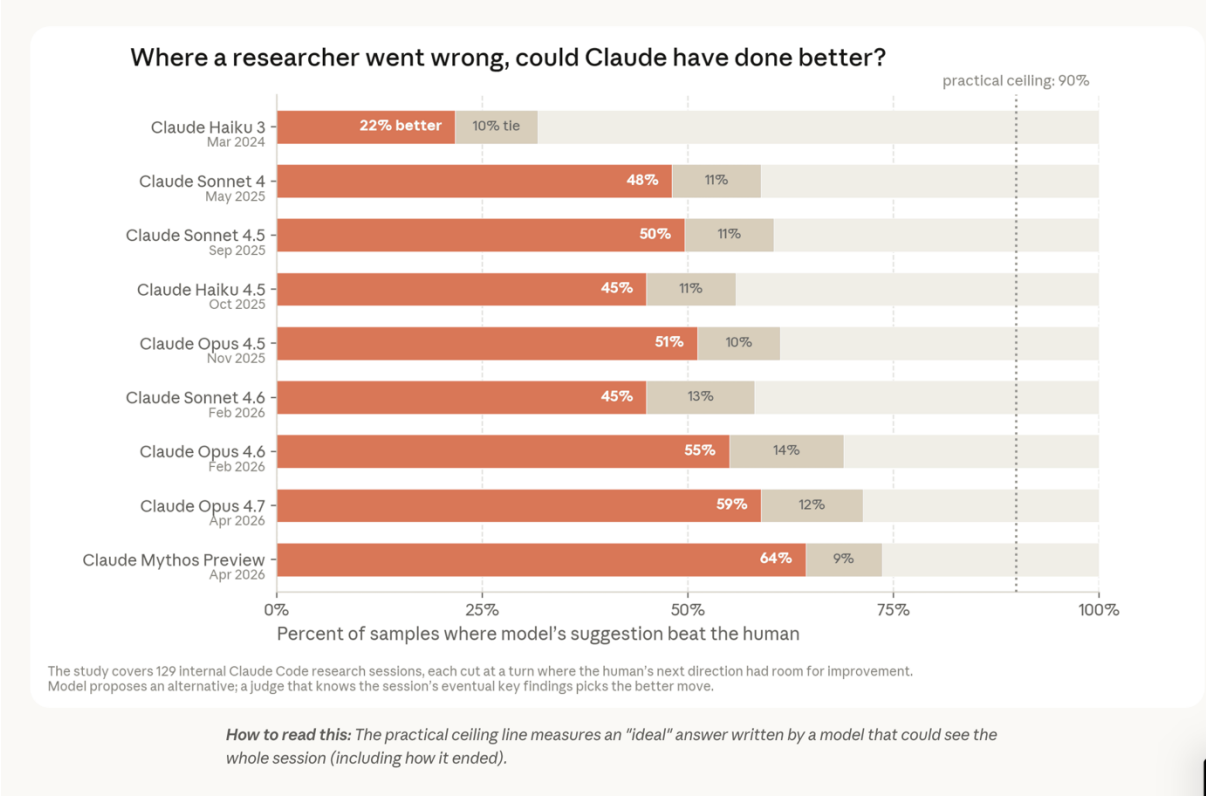




Source: Patwardhan et al. (2025), GDPVal.

Recursive Self Improvement (RSI)

The idea that that sufficiently advanced AI systems can improve themselves without human input. Formal definition contested.



Source: Anthropic Blog. “When AI builds itself.”

Economist Forecasts of GDP

By 2030:

RAPID PROGRESS

Outperforms top humans in research, coding and leadership; award-winning creative works; nearly all physical tasks

RESEARCH
Years of work in days

PROBLEM-SOLVING
Outperforms humans at many jobs

CREATIVITY
Grammy/Pulitzer-caliber media

AGENCY
CEO-level operations

ROBOTICS
Nearly all home and industrial tasks, faster

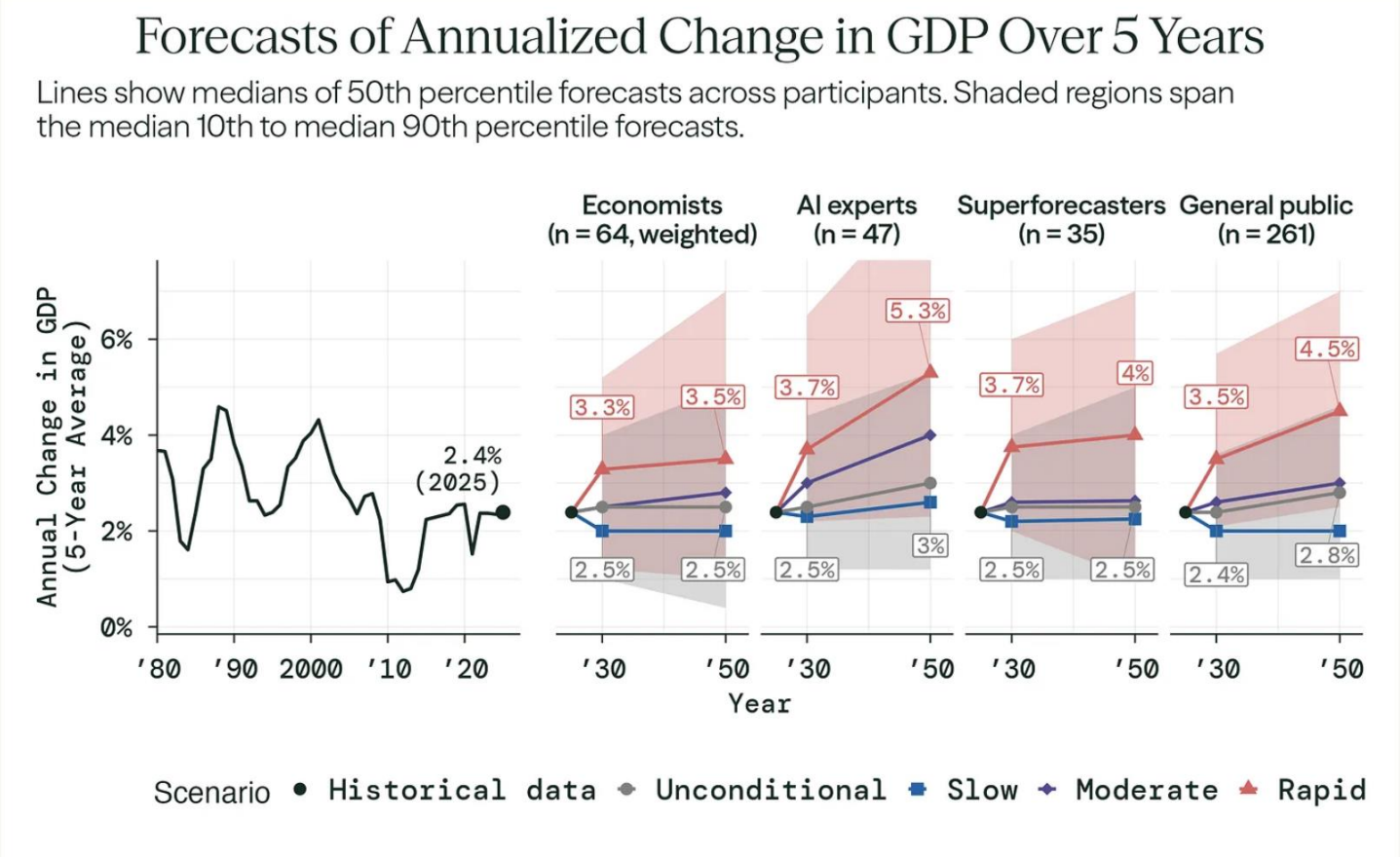


Figure 2: Respondents' median forecasts of the annualized change in U.S. GDP in the five years leading to 2030 and 2050 in unconditional and various AI progress scenarios. The labeled points show median 2030 and 2050 forecasts for the unconditional (gray) and rapid AI progress (red) scenarios.

Source: Murphy et al. (2025). The Longitudinal Expert AI Panel.

Economist Forecasts of Labor Force Participation Rate

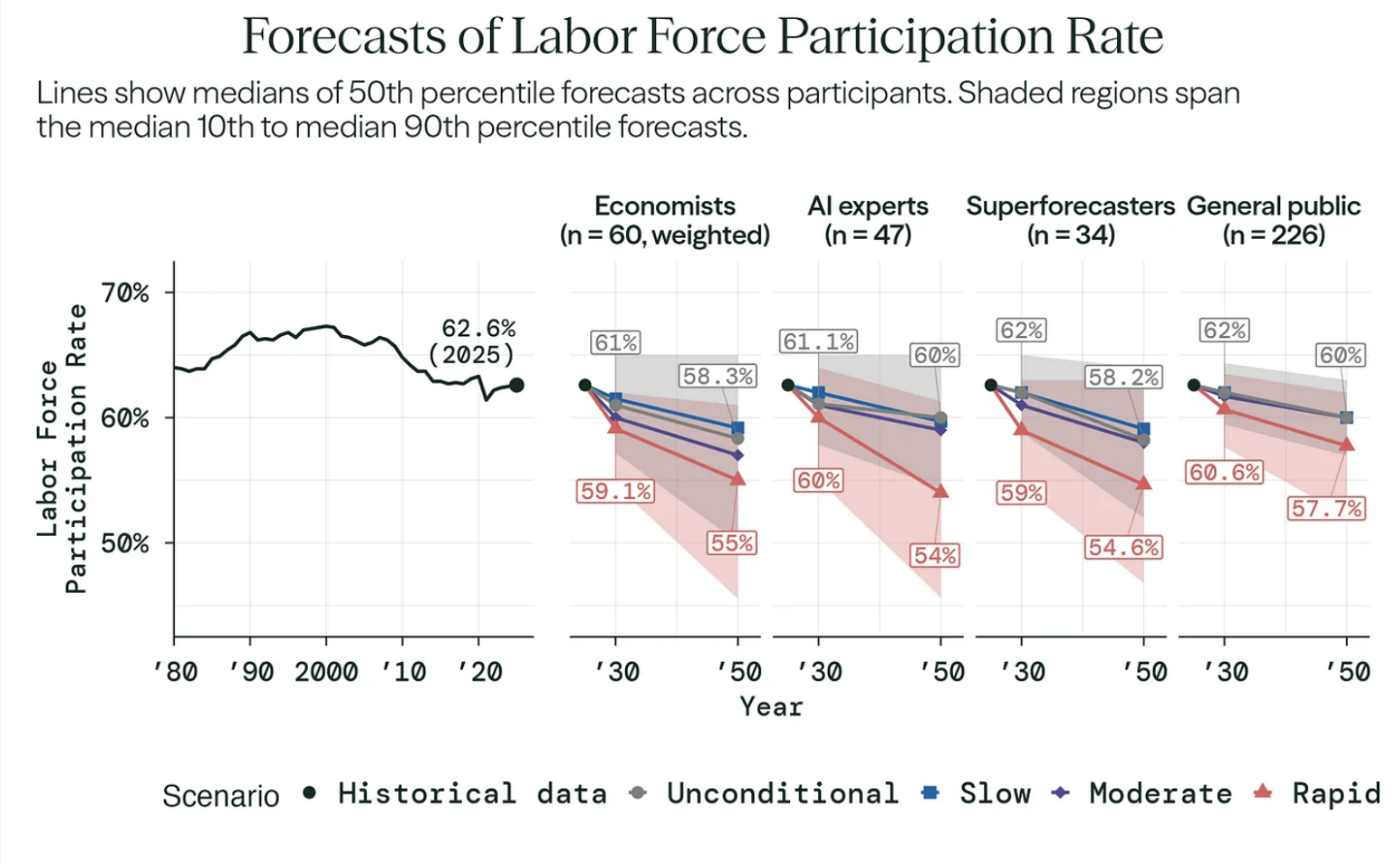


Figure 3: Respondents' median forecasts of U.S. LFPR in 2030 and 2050 in unconditional and various AI progress scenarios. The labeled points show median 2030 and 2050 forecasts for the unconditional (gray) and rapid (red) scenarios.

Source: Murphy et al. (2025). The Longitudinal Expert AI Panel.

Can prediction markets do better?

We need well-capitalized prediction markets for AI impacts

And how to create them

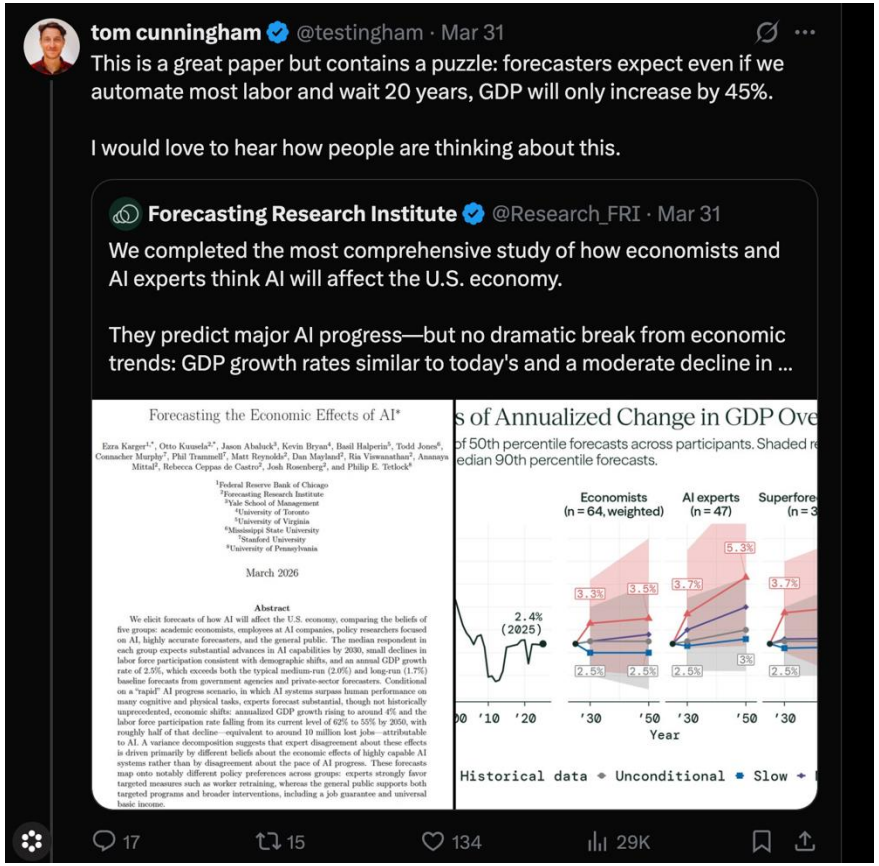
ANDREY FRADKIN, BRIAN JABARIAN, AND ANDREW KOH

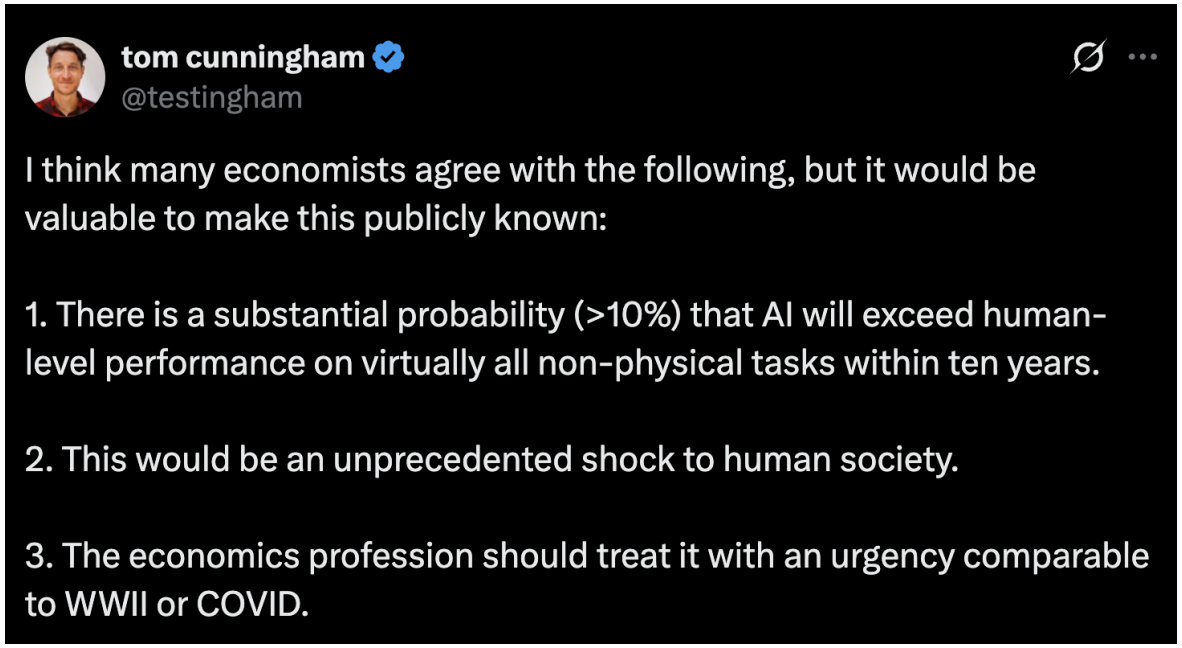
MAY 19, 2026

13 3 6 Share

As our readers have surely noticed, there's enormous disagreement about the likely economic effects of AI. Some AI forecasters, especially those emphasizing recursive self-improvement, expect AI capabilities to soon surpass that of most human workers. As a result, they expect large-scale labor market displacement, beginning with white-collar work and spreading to all of labor in the long run. The AI 2027 scenario lays out a year-by-year path to transformative AI and implies large-scale labor displacement. Other economists, in contrast, are more skeptical of such extreme rapid take-offs, emphasizing adoption frictions, organizational bottlenecks, complementary investments, and the slow diffusion of new production processes.

The Wisdom of Tom Cunningham





The market for intelligence

Motivation

- Most existing research focuses on:
  - Productivity
  - Employment
  - Macroeconomic implications
  - As a tool for businesses
- Today: **industrial organization + market design** of the AI industry

Several Papers:

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

Mert Demirer† Andrey Fradkin‡ Nadav Tadelis§

March 22, 2026

Accepted at the Journal of Economic Perspectives.

How Firms Use LLMs: Adoption Patterns and Market Heterogeneity *

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

March 30, 2026

+ Ongoing work with Mert and Aaron Kaye on estimating demand.

Why focus on the industrial organization of AI?

(We don't answer all of these!)

Demand Side:

- Vertical and horizontal differentiation, what is model 'quality'?
- What is the price elasticity?
- How quickly do better models diffuse?

Supply Side:

- How do firms compete with each other?
- What is the role of open source?
- What are the private and/or social returns? Anti-trust / regulation?

Most of the results today are descriptive

Use two main data sources: OpenRouter and Microsoft Azure Foundry, with coverage between 2022 and **end of 2025**.

Document:

- Models, creators, and providers.
- Pricing trends, and prices per unit of intelligence.
- Diffusion, market shares, and concentration. Also, by use case and industry.
- Event studies of market entry.

Main Facts

1. Rapid growth of LLMs and inference providers.

2. Dramatic fall in prices (1000x) in two years.

3. Dynamism in market leadership.

4. Horizontal and vertical differentiation.

5. Demand for tokens. Elasticities imply no short-run Jevons' effects.

6. Limited multi-homing across models.

Related work

- Occupational exposure (Bryjnolfsson et al. 2018, Eloundou et al. 2024, Handa et al., 2025).
- Macro modeling (Acemoglu and Restrepo (many), Autor and Thompson, 2025, Jones et al. (many))
- RCTs (Dell'Acqua et al. (2026), Noy and Zhang (2023))
- Preferences over text (huge CS literature)
- Open source models: Nagle and Yue (2025)

Presentation Roadmap

1. Introduction and Motivation
2. **Data Sources and Institutional Background**
3. Supply and Pricing
5. Demand
6. Substitution and Market Expansion
7. Usage and Heterogeneity
8. Conclusion

Types of LLM Markets

Packaged / Consumer LLMs:

- Individual use, freemium
- ChatGPT, Gemini, personal Codex / Claude Code

LLM API Market (This Paper):

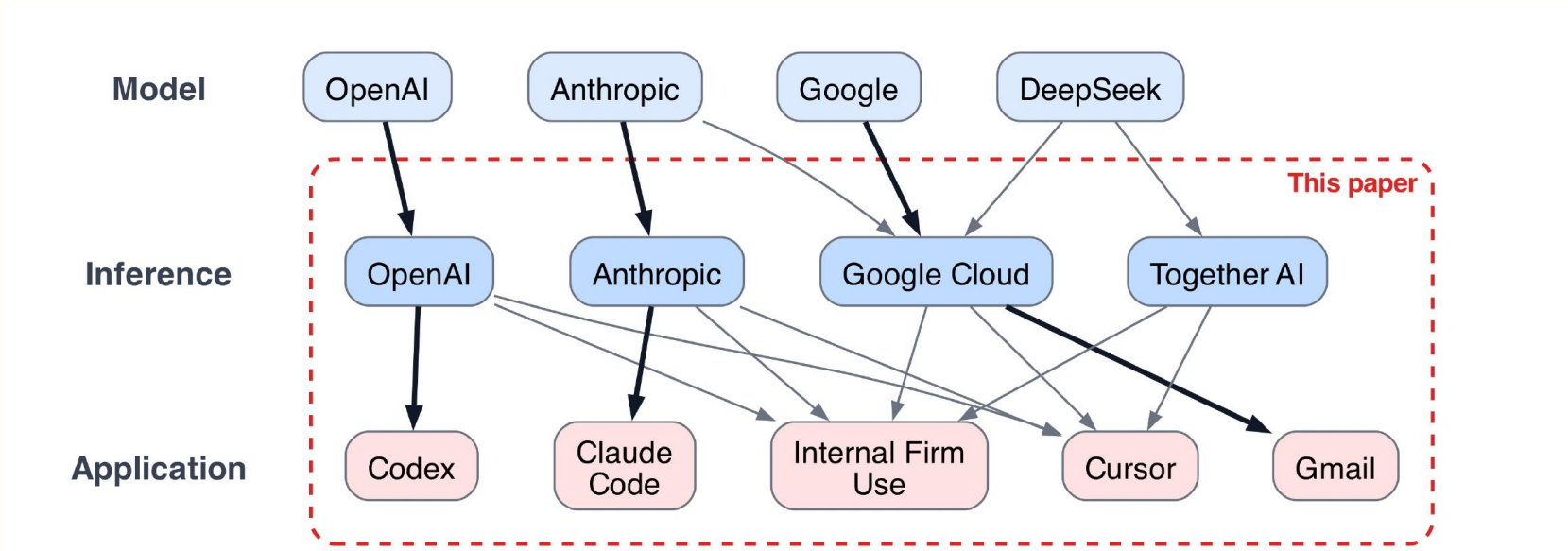
- API calls and usage based pricing.
- AWS, Azure, OpenAI, Together AI, etc.

Custom Enterprise LLMs

- Vertical specific models, Customized Pricing
- Consulting Firms, OpenAI, Sierra, Harvey, etc...

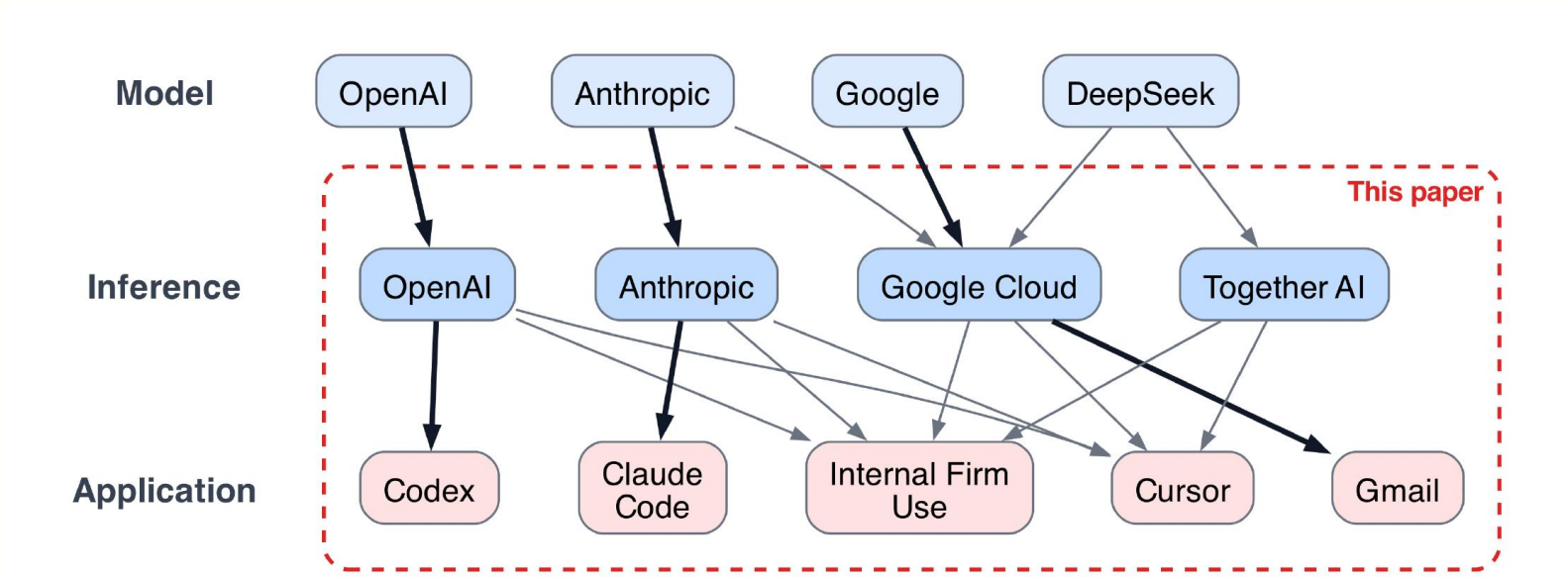
LLM Creators

- Develop model architecture.
- Assemble data.
- Conduct training runs.



LLM Inference Providers

- Big 3: AWS, Azure, and Google Cloud Platform.
- OpenAI and Anthropic run their own endpoints.
- Together AI, Cerebras, Groq, and others specialists.



LLM API Pricing

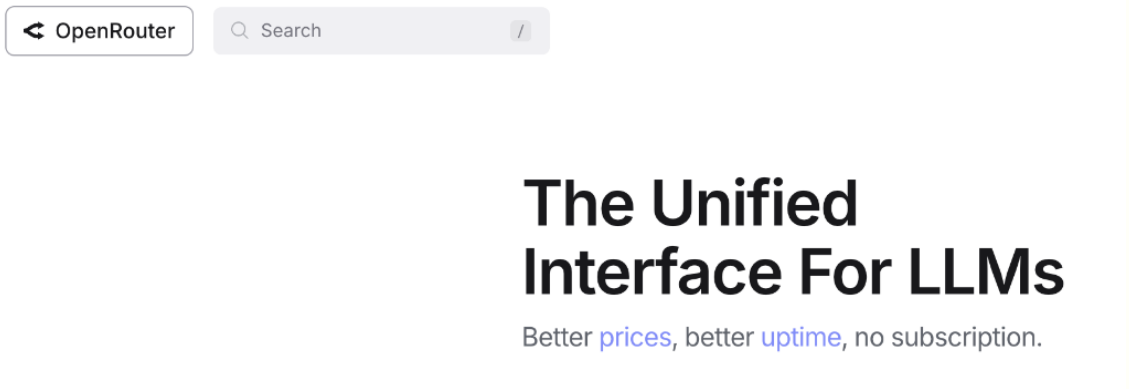
Posted prices:

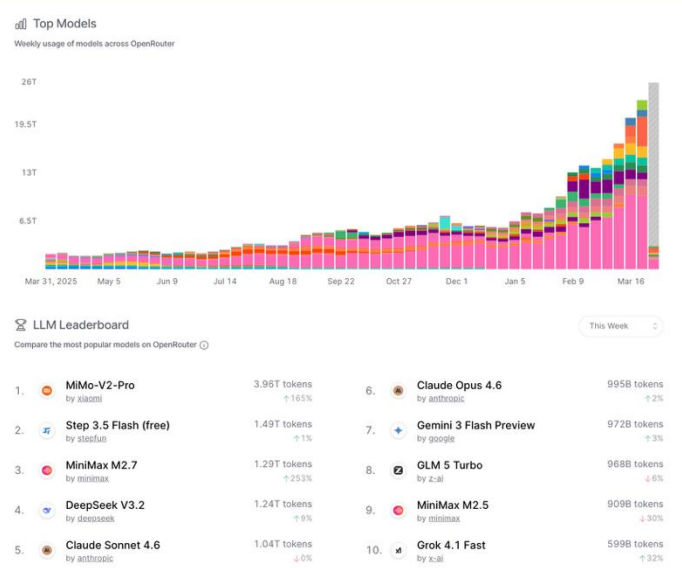
- Cache tokens
- Input tokens
- Output tokens
- Reasoning tokens

LLM Aggregators and OpenRouter

LLM aggregators are platforms that sit between developers and model providers.

- They lower switching costs and offer other conveniences.
- OpenRouter is the biggest one, serving over 25 trillion tokens per week.
- Pricing: 5% to 5.5%





OpenRouter Data

- Model-level data, includes every provider, usage, prices, and other characteristics.
- Category-level data.
- App level data (e.g, Cline, Roo Code, LiteLLM in our sample). Today: OpenClaw, Hermes

Some data available in July 2023, comprehensive data after April 2025.

Side note: We now have underlying OpenRouter data, but not for today. Stay tuned!

Azure Foundry Data

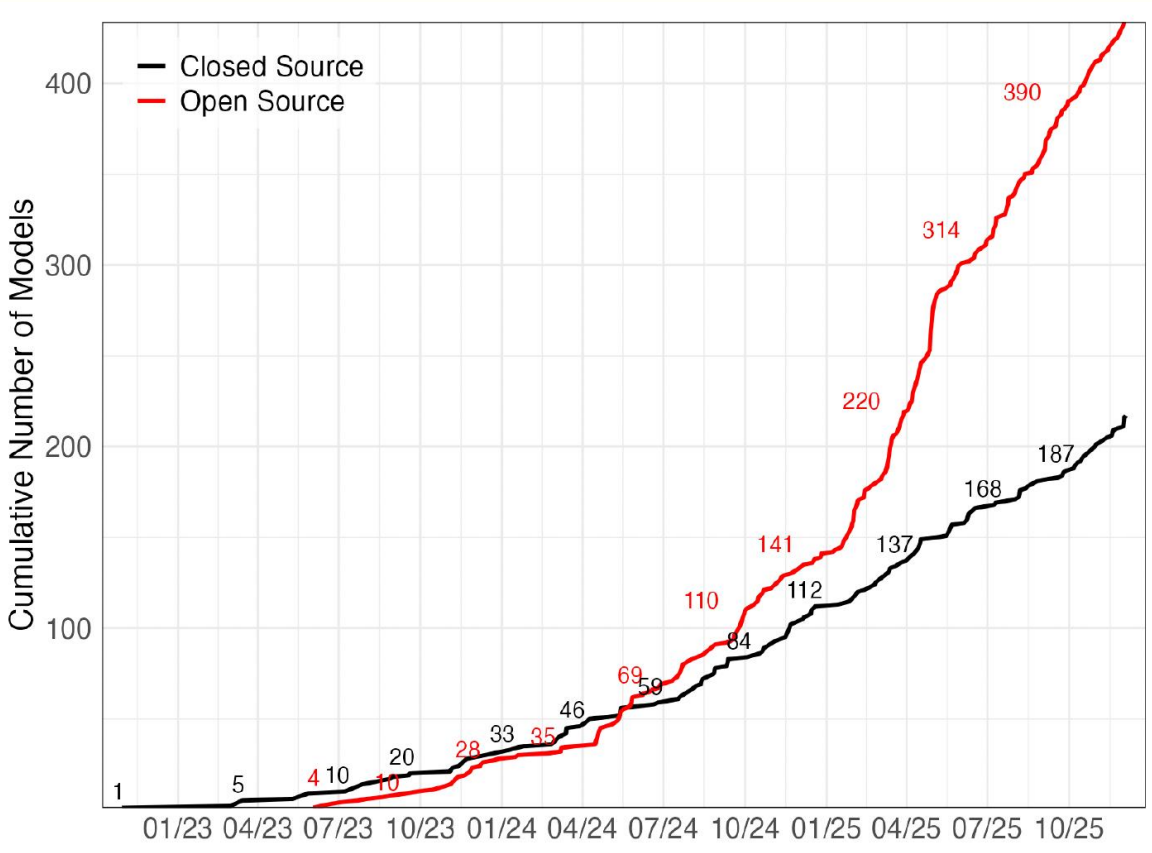
- Foundry allows enterprises to access models hosted by Microsoft through an API.
- Serves OpenAI and xAI models, and recently Anthropic models. Not Google models. Also serves open-source models.
- Enterprise-day level usage + enterprise industry.

We did lots of data cleaning to sync OpenRouter, Azure, and other data.

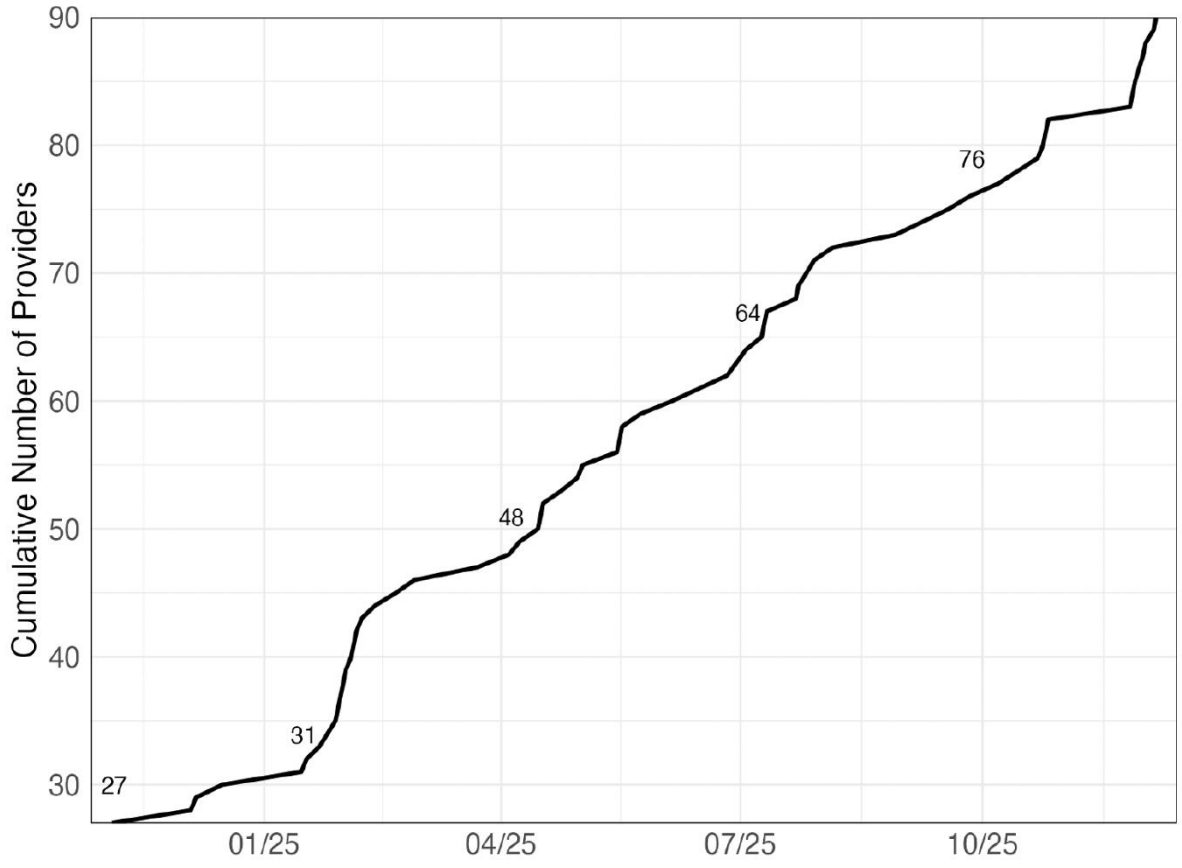
Presentation Roadmap

1. Introduction and Motivation
2. Data Sources and Institutional Background
3. **Supply and Pricing**
5. Demand
6. Substitution and Market Expansion
7. Usage and Heterogeneity
8. Conclusion

Model and Creator Growth

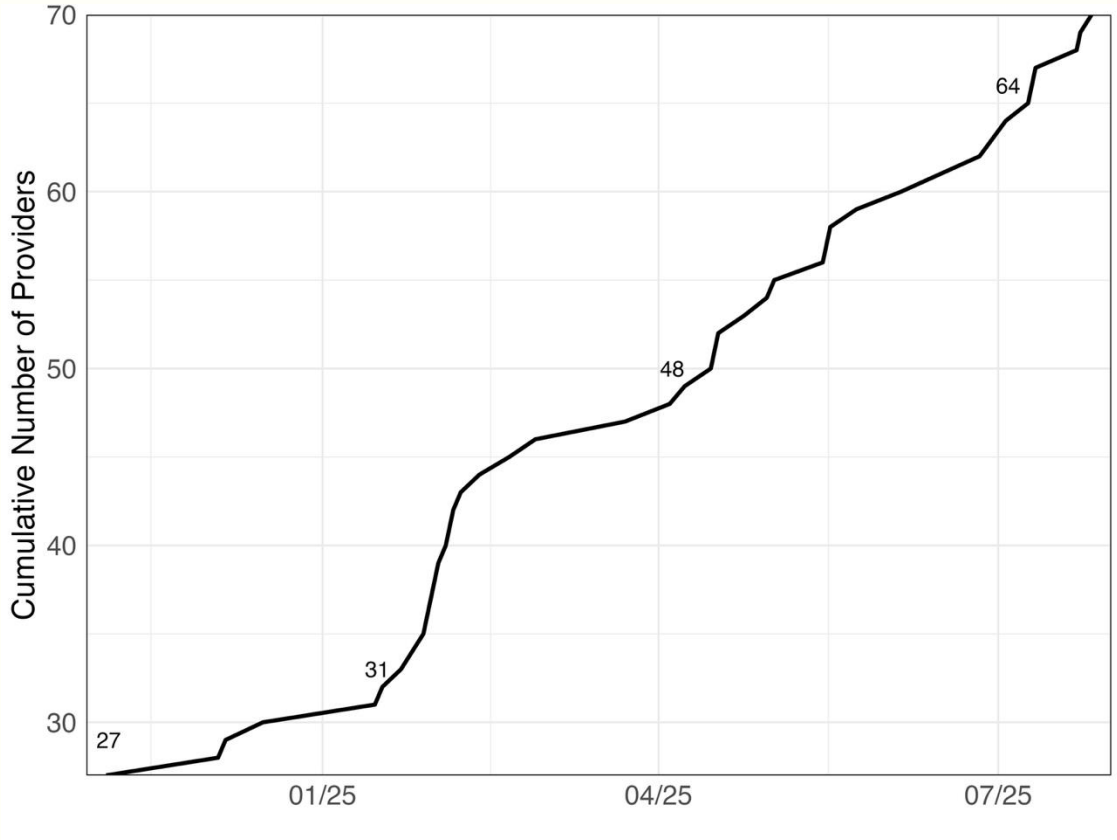


Number of Models

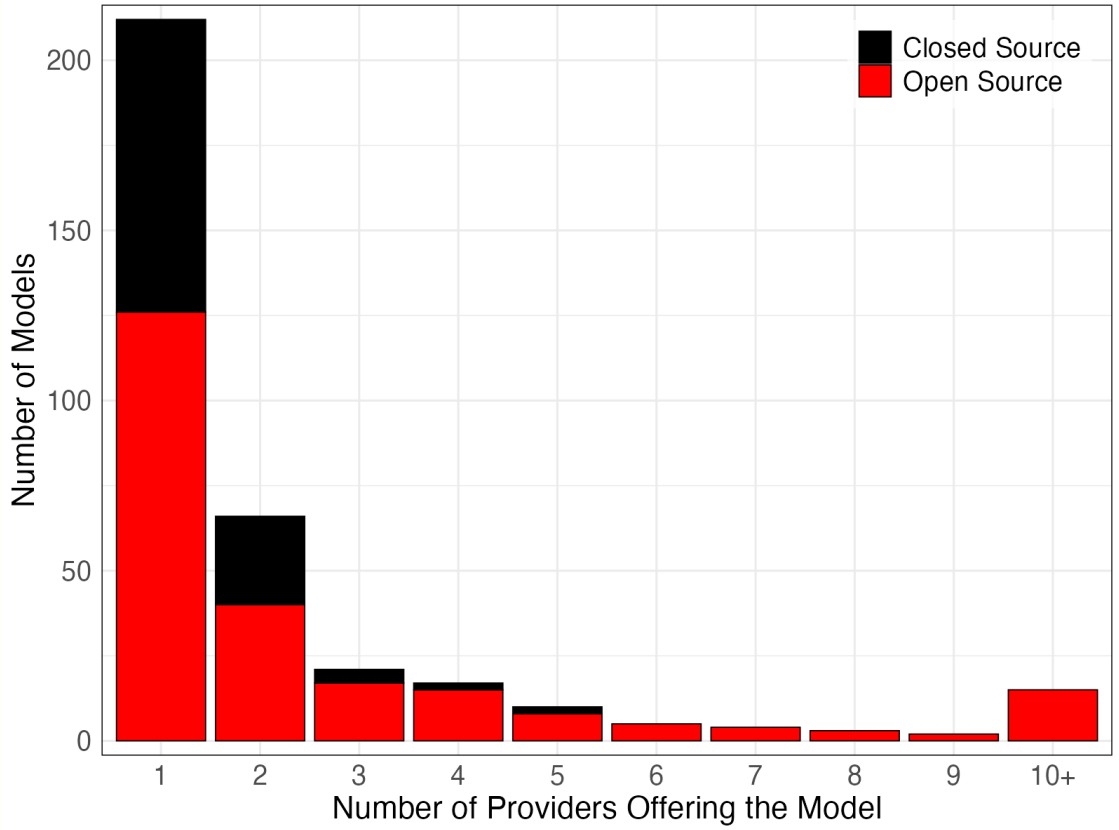


Number of Creators

Providers

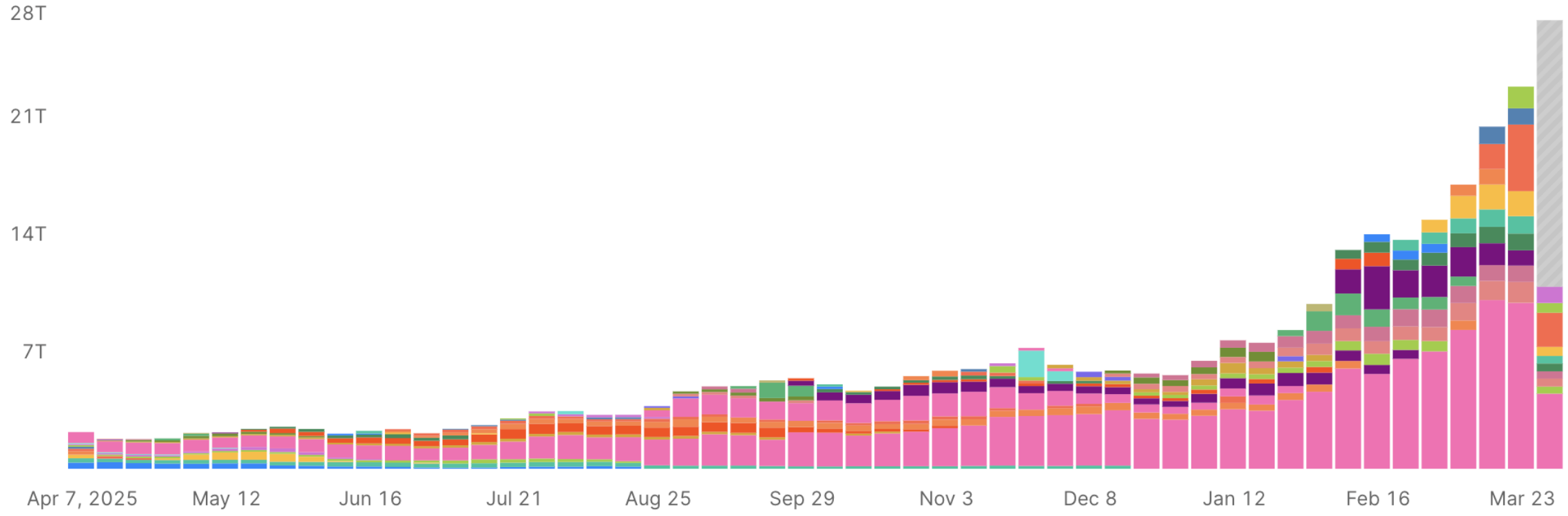


Number of Providers



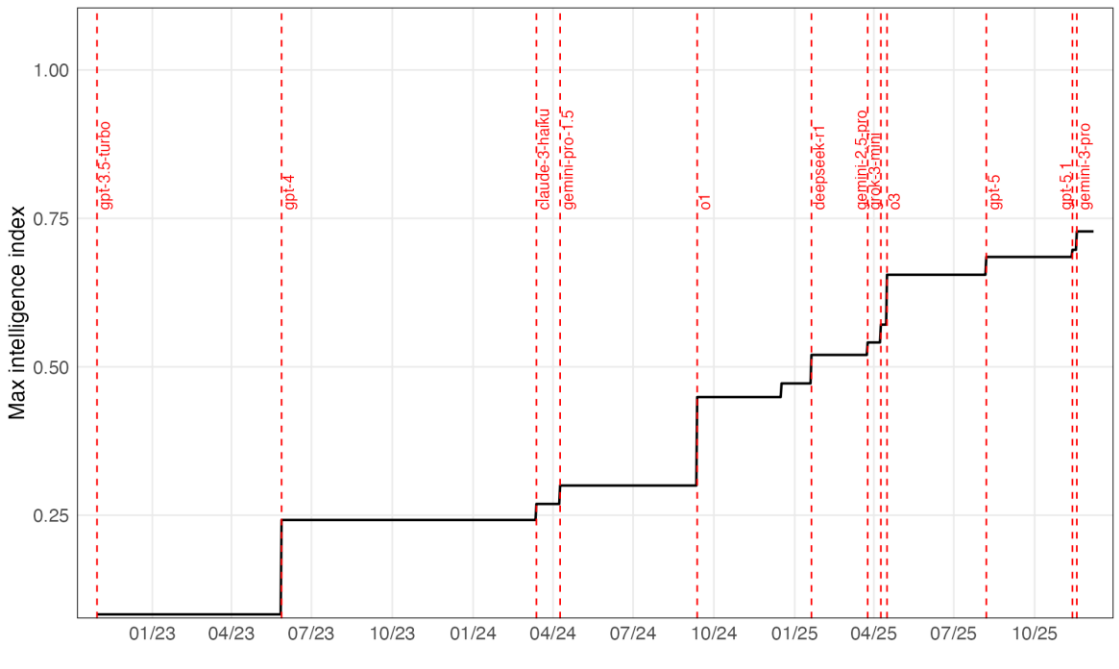
Providers per Model

Trillions of total tokens demanded (OpenRouter)

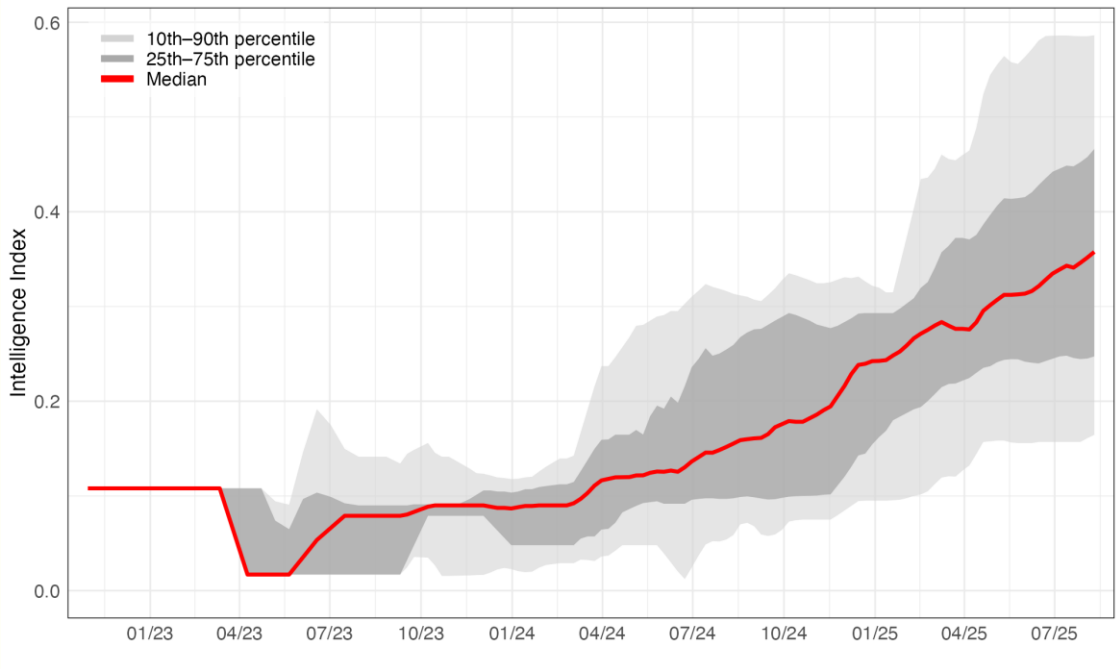


Measuring Model Intelligence

- Lots of benchmarks: MMLU Pro, GPQA, HLE, LiveCodeBench, etc....
- Model performance across these is very correlated.
- We use an index from Artificial Analysis (0 to 100).

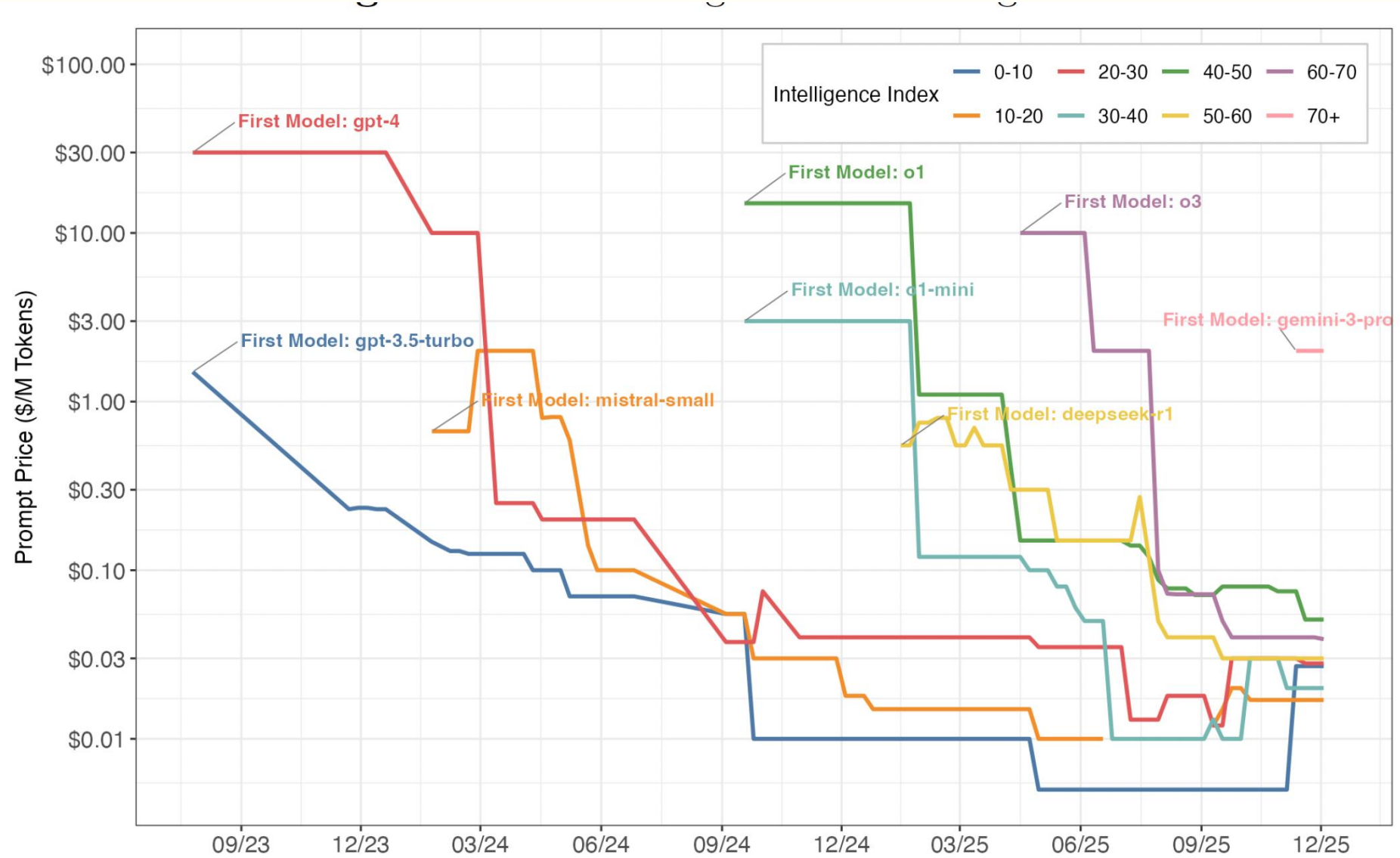


Most Intelligent Models



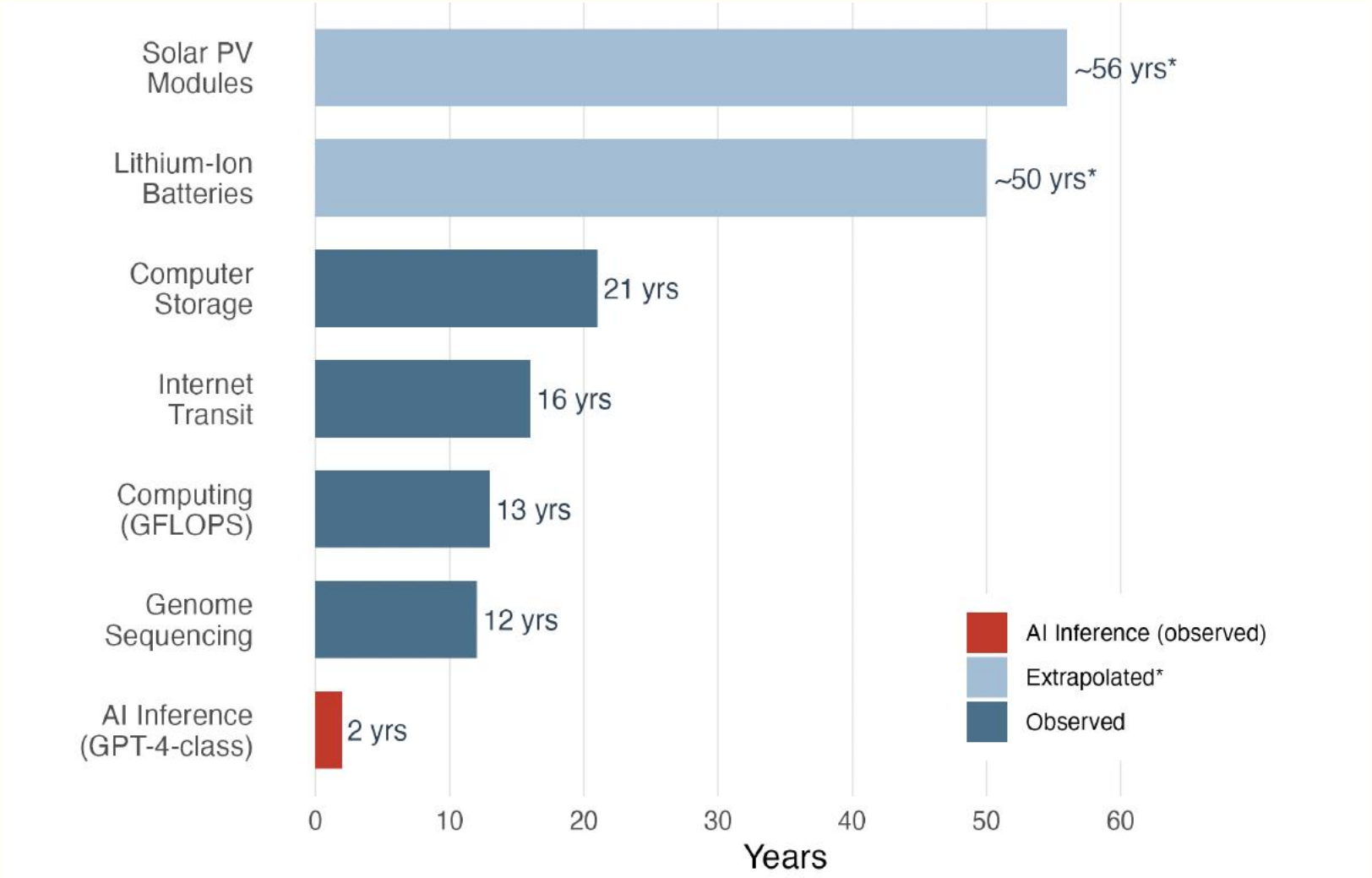
Intelligence of Recent Models

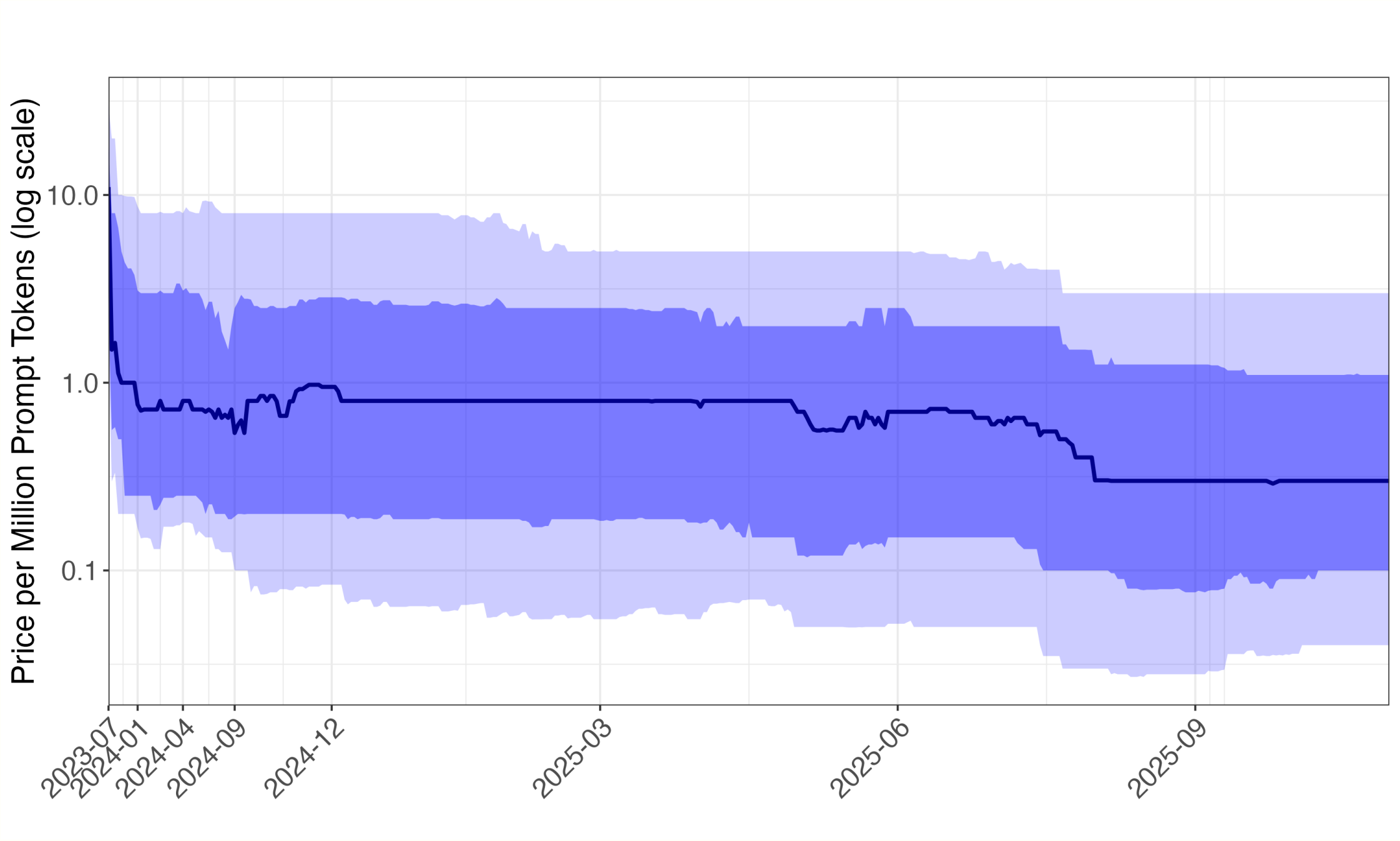
Pricing



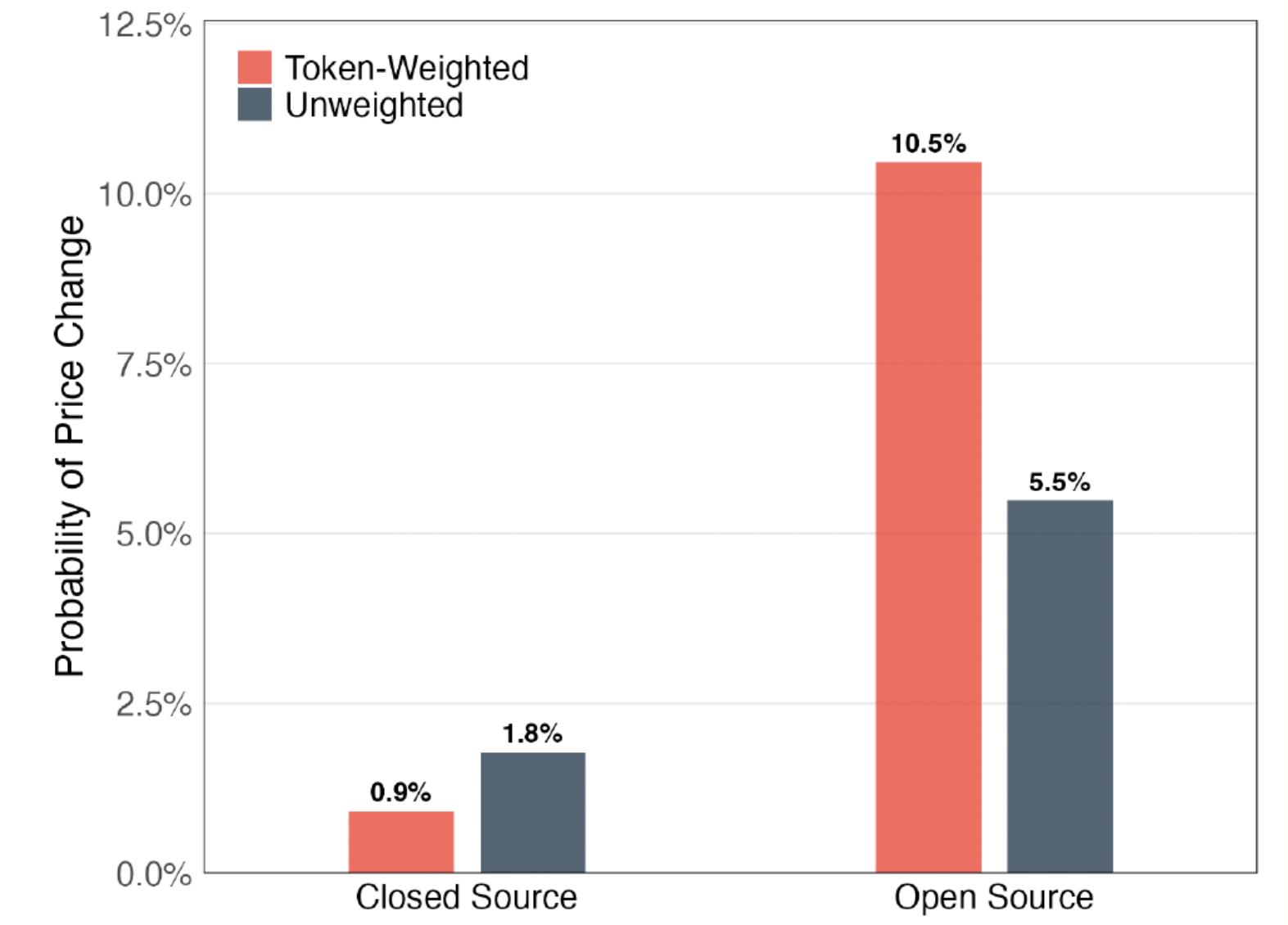
Minimum price per million tokens

Time to 1000x drop in prices across major technologies.

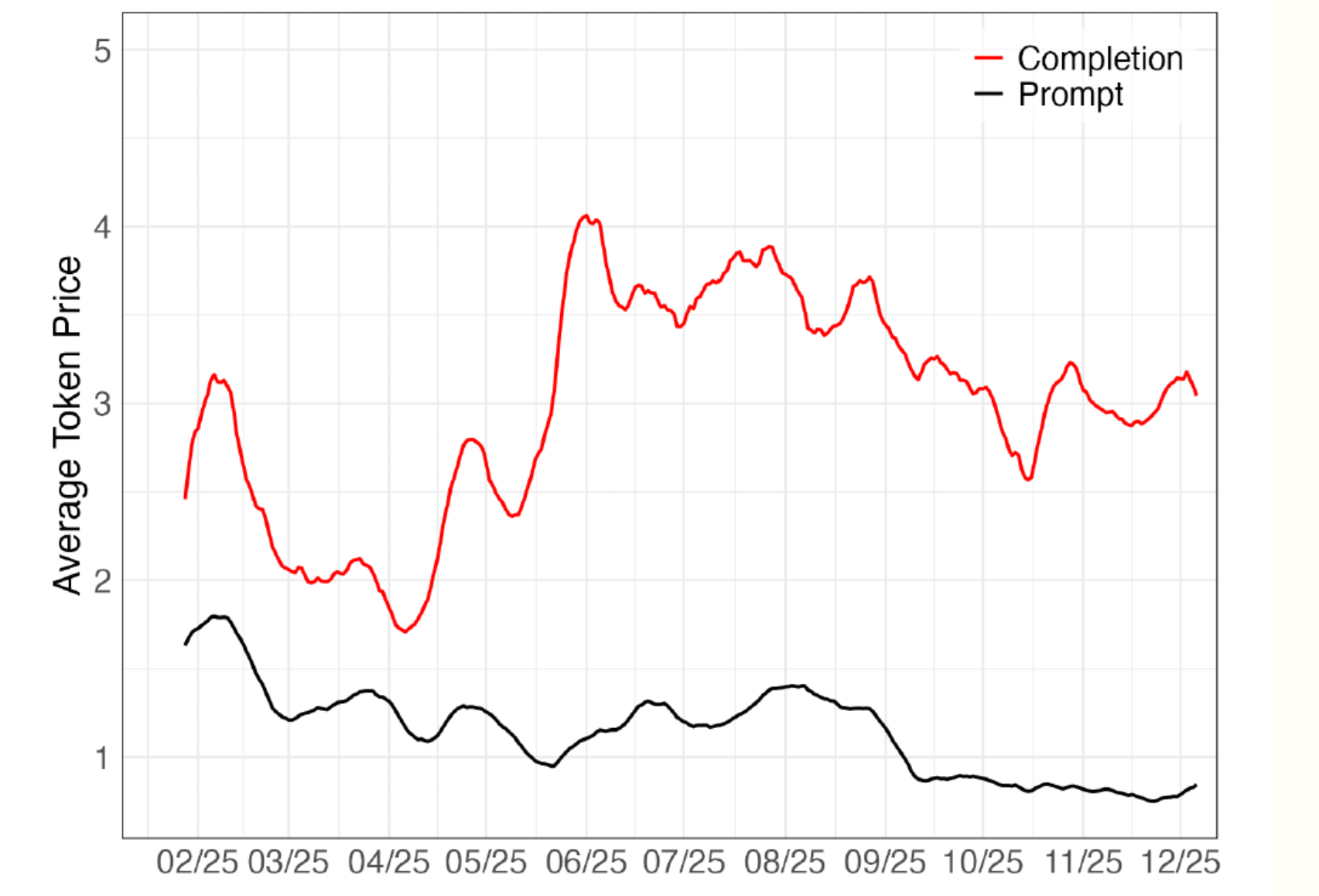




Price distribution of recently released models

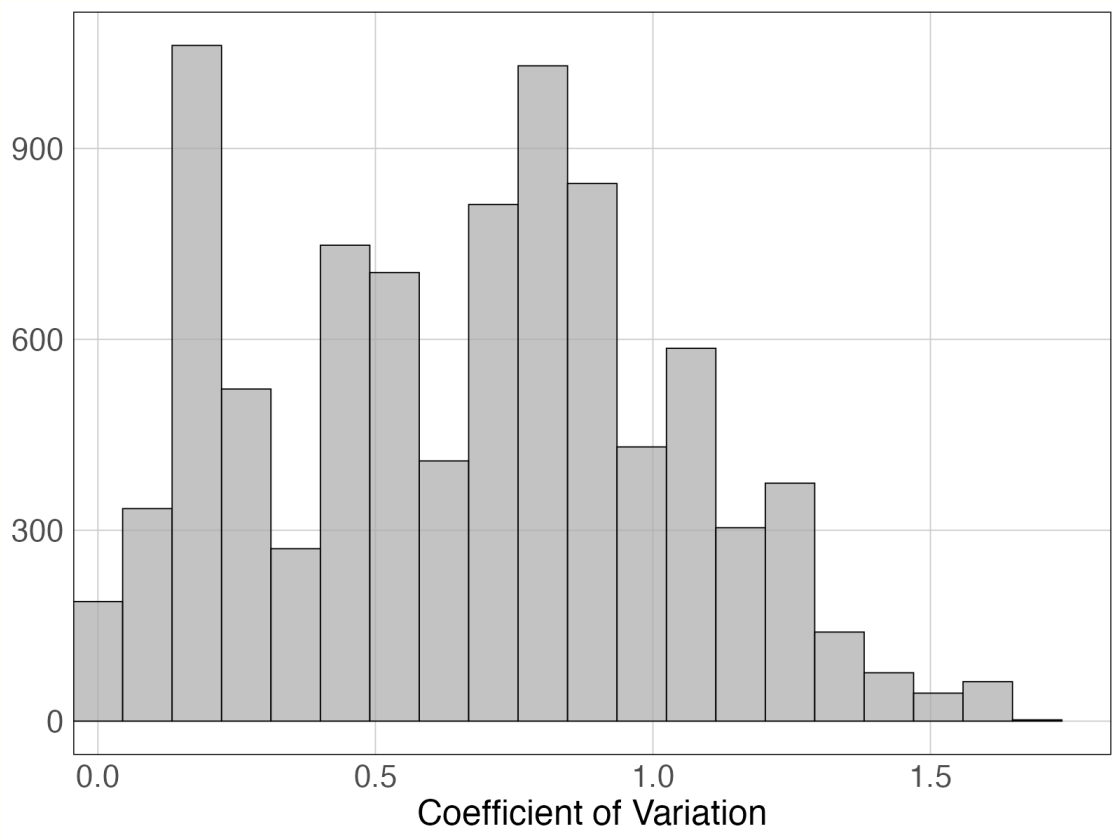


Open-source models experience more frequent price changes

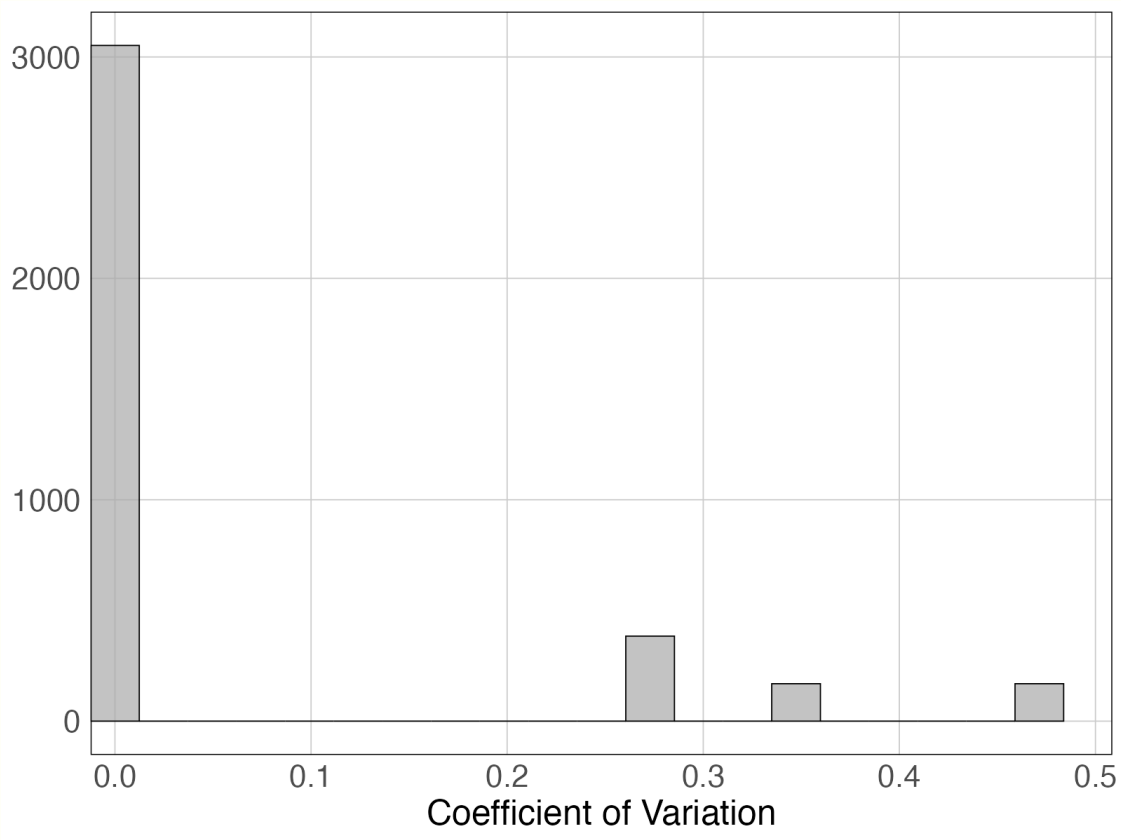


Usage-weighted price paid per token decreasing

Variation in prices within model

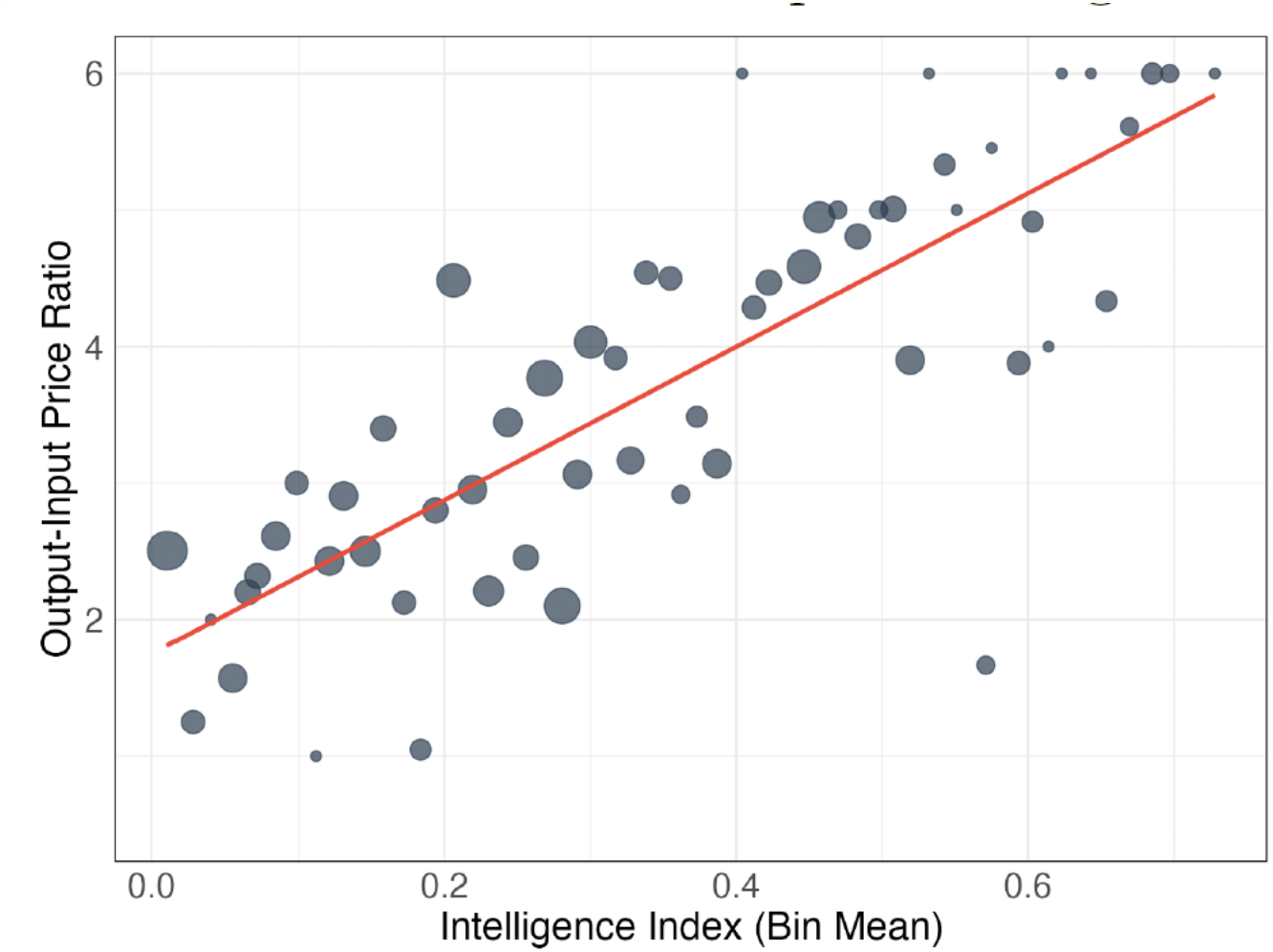


Open-source



Closed-source

Intelligent models have more expensive output tokens



Hedonic Price Regression: More intelligent models are more expensive

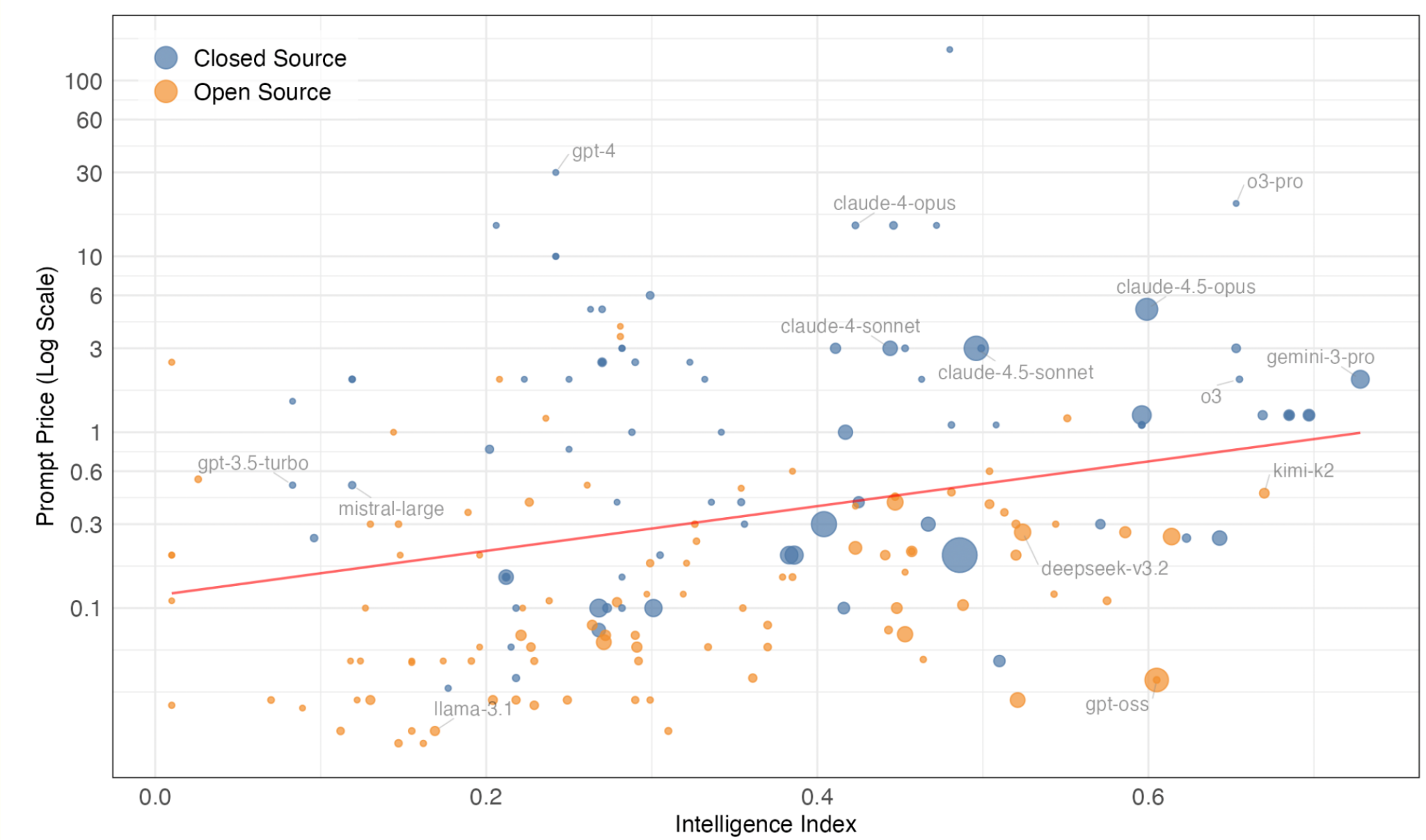
	Log Price per Million Prompt Tokens				
	(1)	(2)	(3)	(4)	(5)
Intelligence Index	0.039***	0.029***	0.040***	0.052***	0.047***
	(0.009)	(0.010)	(0.011)	(0.014)	(0.017)
Open Source		-2.46***	-2.17***	-2.00***	-1.91***
		(0.266)	(0.267)	(0.458)	(0.440)
Supports Reasoning		-0.107	0.132	-0.146	0.097
		(0.307)	(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				✓	✓

Open source models are vastly cheaper

	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				✓	✓

Presentation Roadmap

1. Introduction and Motivation
2. Data Sources and Institutional Background
3. Supply
5. **Demand**
6. Substitution and Market Expansion
7. Usage and Heterogeneity
8. Conclusion



Usage by model intelligence and price

Motivation for estimating demand: Jevons' Paradox

- If technological improvements lower the cost of using a resource, does demand increase so much that the total resource consumption goes up? Key for thinking about investments in compute.
- Formal condition on the aggregate elasticity of demand, namely that it is greater than 1.
- Today: elasticity across providers / models. Aggregate elasticity is strictly smaller in magnitude, so if little elasticity across models / providers, even less response at the aggregate level.

Empirical strategy

- Observation: model, provider, date
- Outcome: Quantity demanded
- Characteristics: Price, latency, throughput, context length.
- Identification here is hard. This is a first attempt.

$$\log(Q_{imt}) = \beta_1 \log(\text{Price}_{imt}) + \beta_2 \log(\text{Throughput}_{imt}) + \beta_3 \log(\text{Latency}_{imt}) + \beta_4 \log(\text{Context}_{imt}) + \gamma_t m + \theta_i m + \varepsilon_{imt}$$

No short-run Jevons' paradox

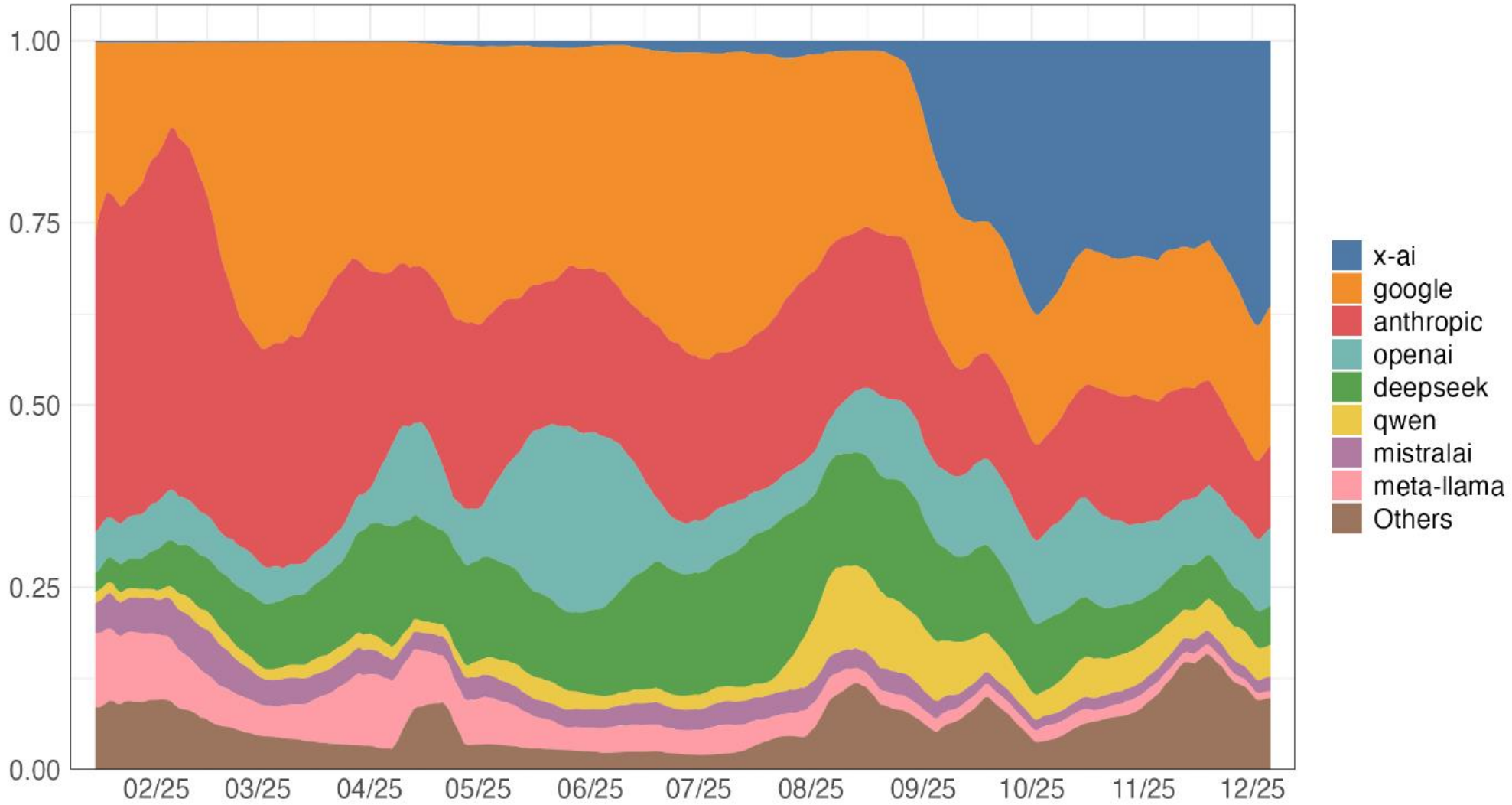
	Log(Daily Tokens)		
	(1)	(2)	(3)
Log(Price)	-0.55***	-1.08***	-1.11***
	(0.09)	(0.19)	(0.22)
Log(Throughput)	-0.40**	-0.01	-0.07
	(0.16)	(0.08)	(0.05)
Log(Latency)	-0.11	-0.31***	-0.13***
	(0.24)	(0.08)	(0.04)
Log(Context Length)	0.82***	0.27***	0.22
	(0.13)	(0.10)	(0.24)
Observations	32,539	32,539	32,539
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			✓

Implications + Ongoing Work

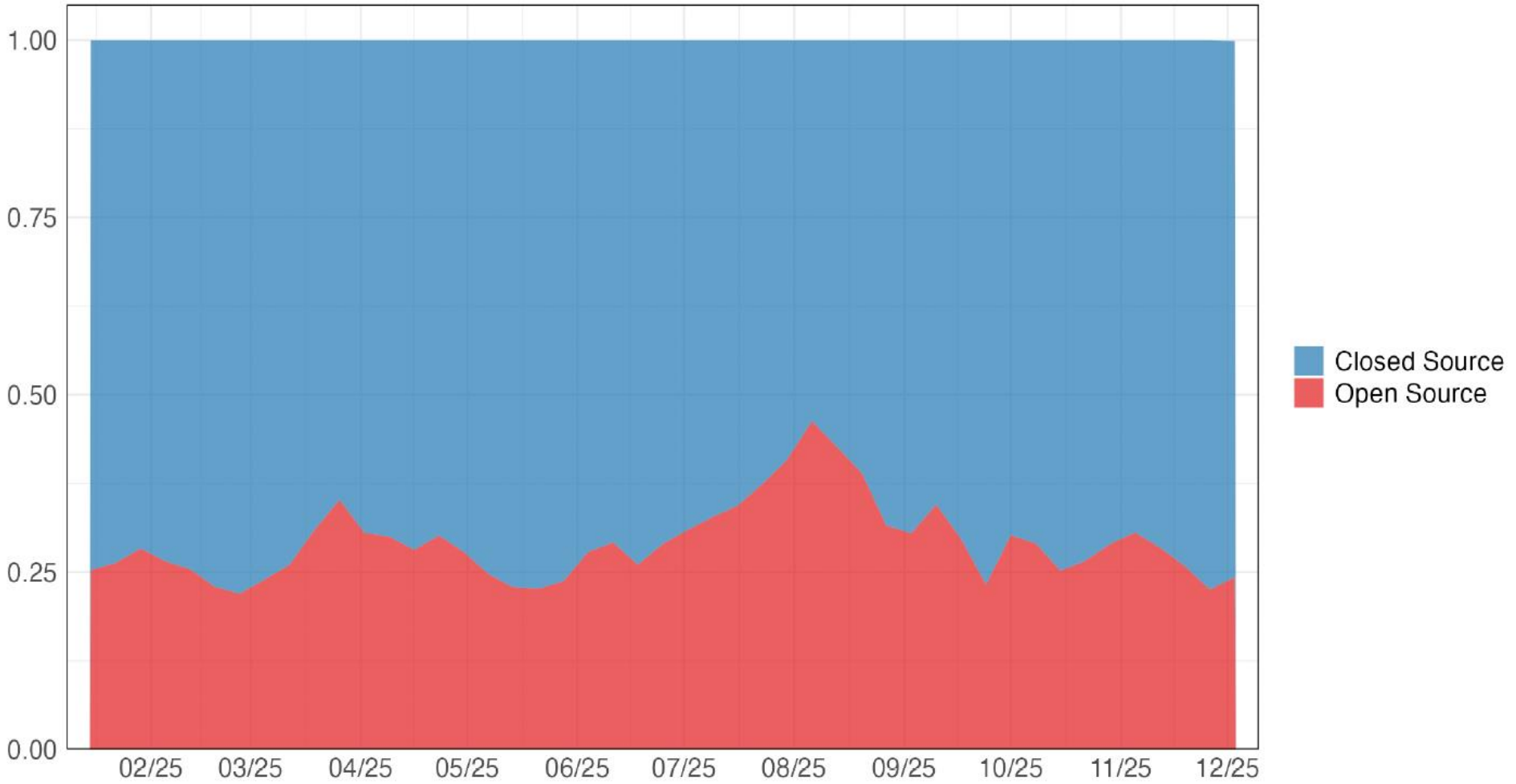
- We are writing a separate demand estimation paper.
- Based on those results, Jevons' paradox is not happening.
- Better models, rather than cheaper models are driving massive usage spikes.
    - See: Sonnet 4.5, Opus 4.5, and GPT 5.4

Evolving Competition and Leadership

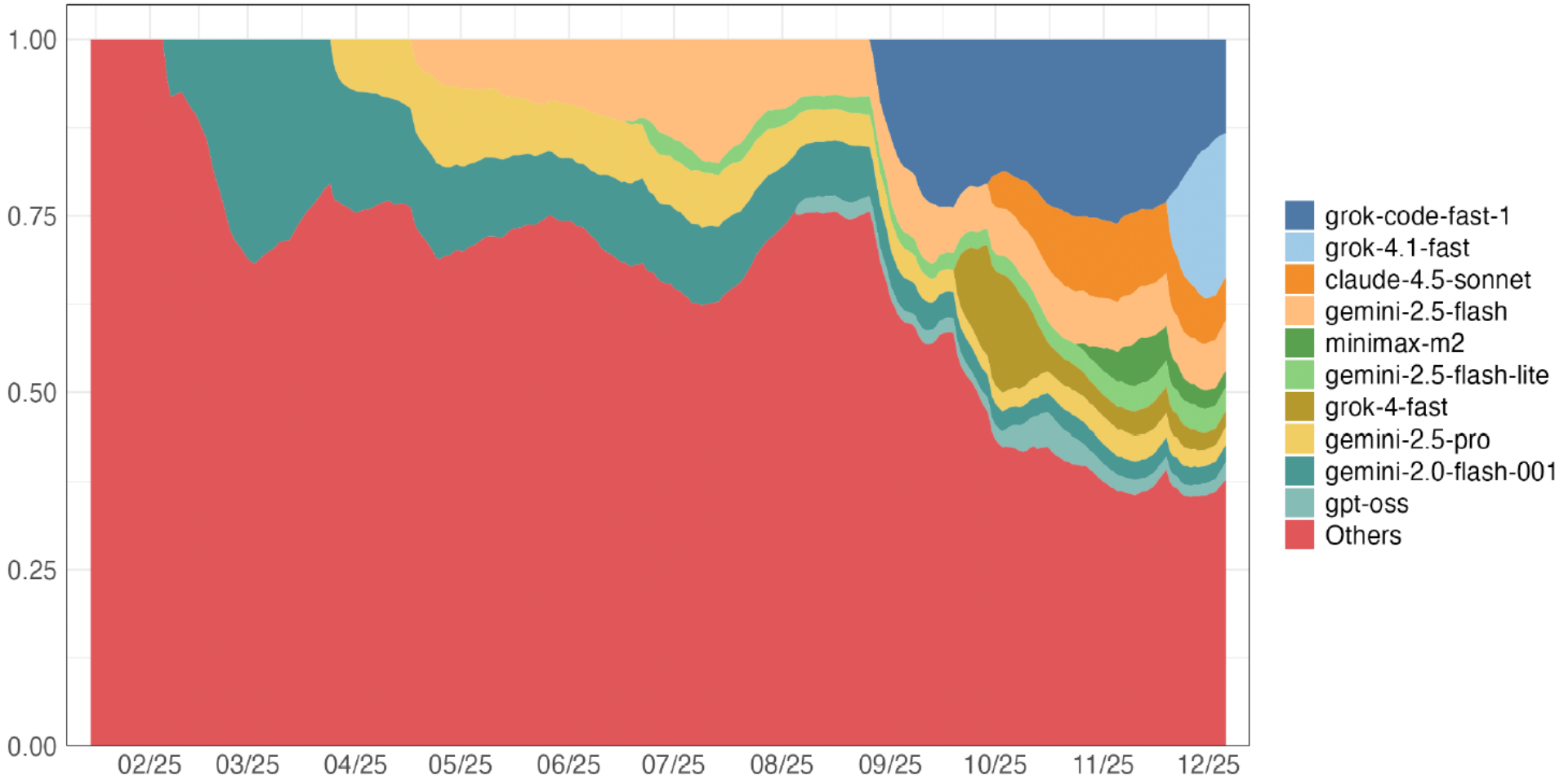
Creator shares over time



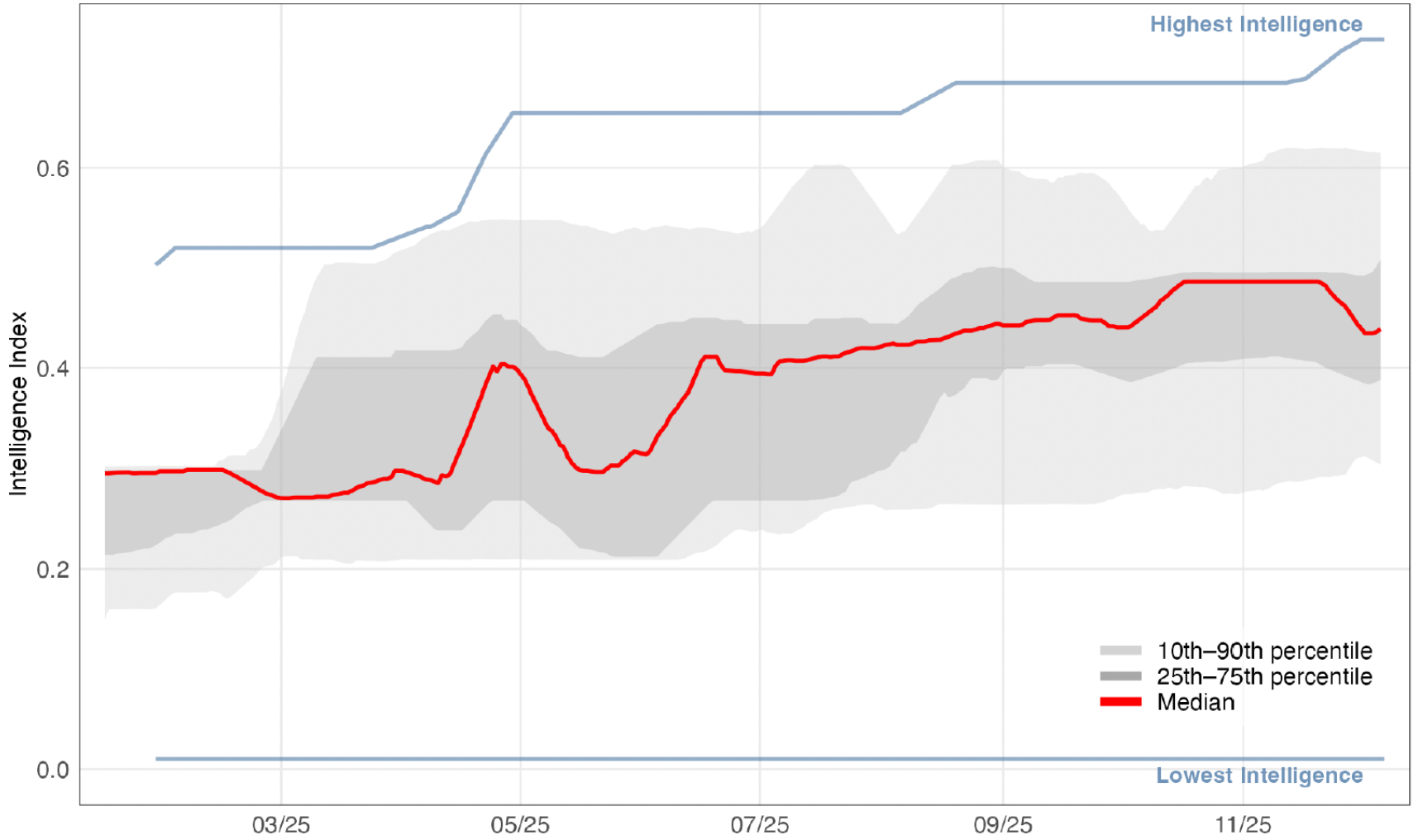
Open source vs closed source over time (Through 2025)



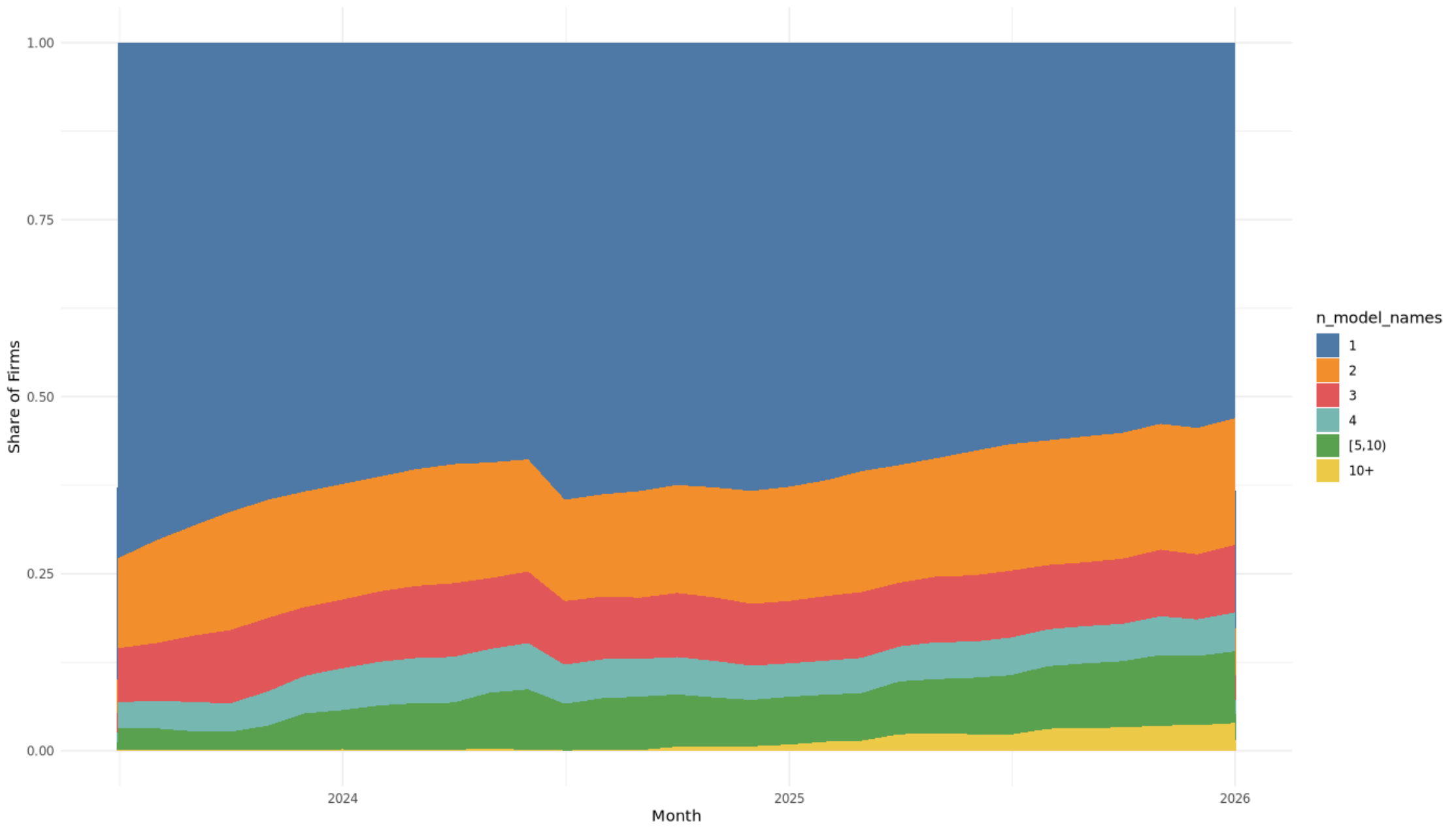
Token share of top 10 models as of end of 2025



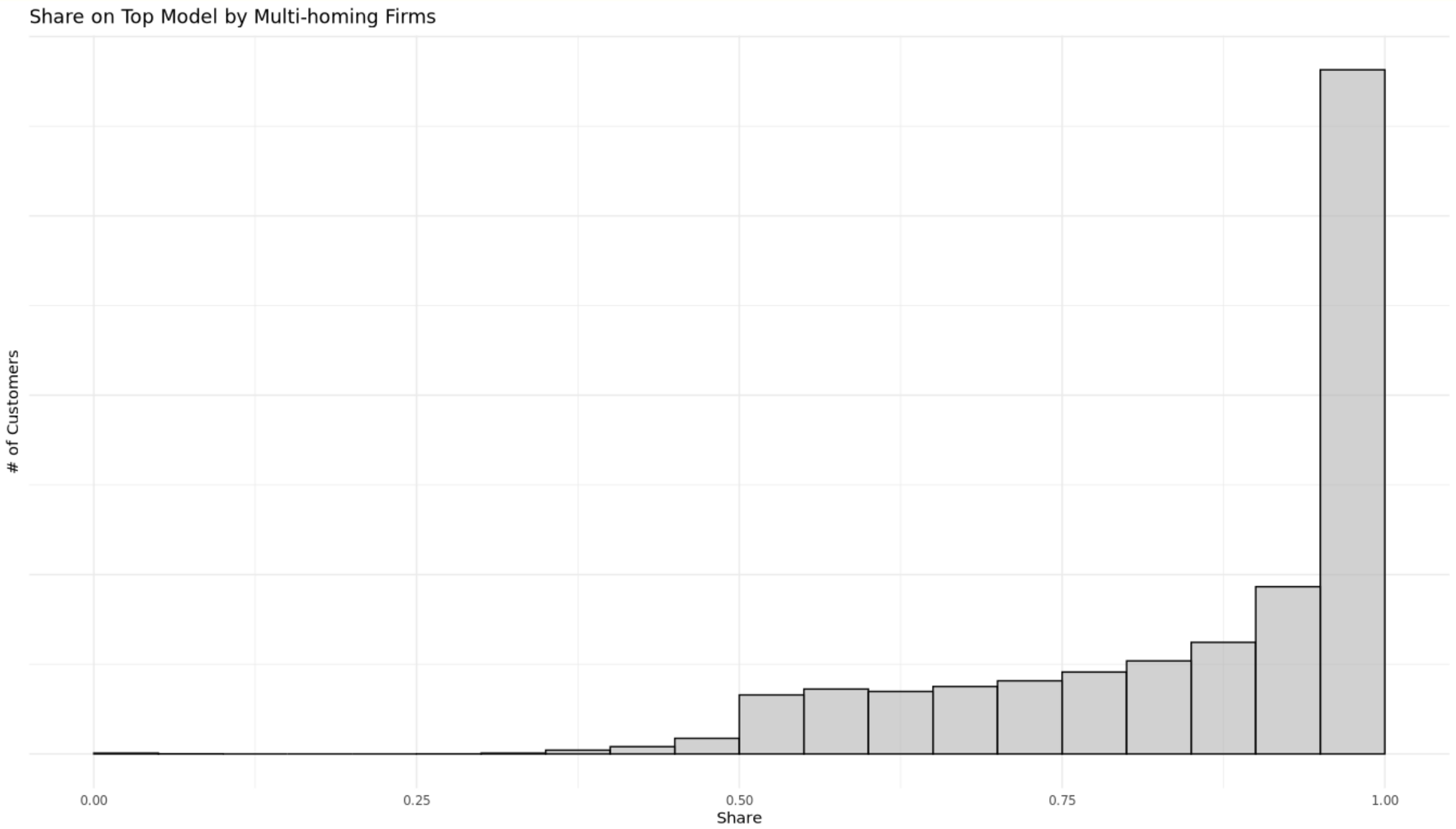
Typical consumed tokens are far from the frontier



Multi-homing



Most firms single-home (Azure)

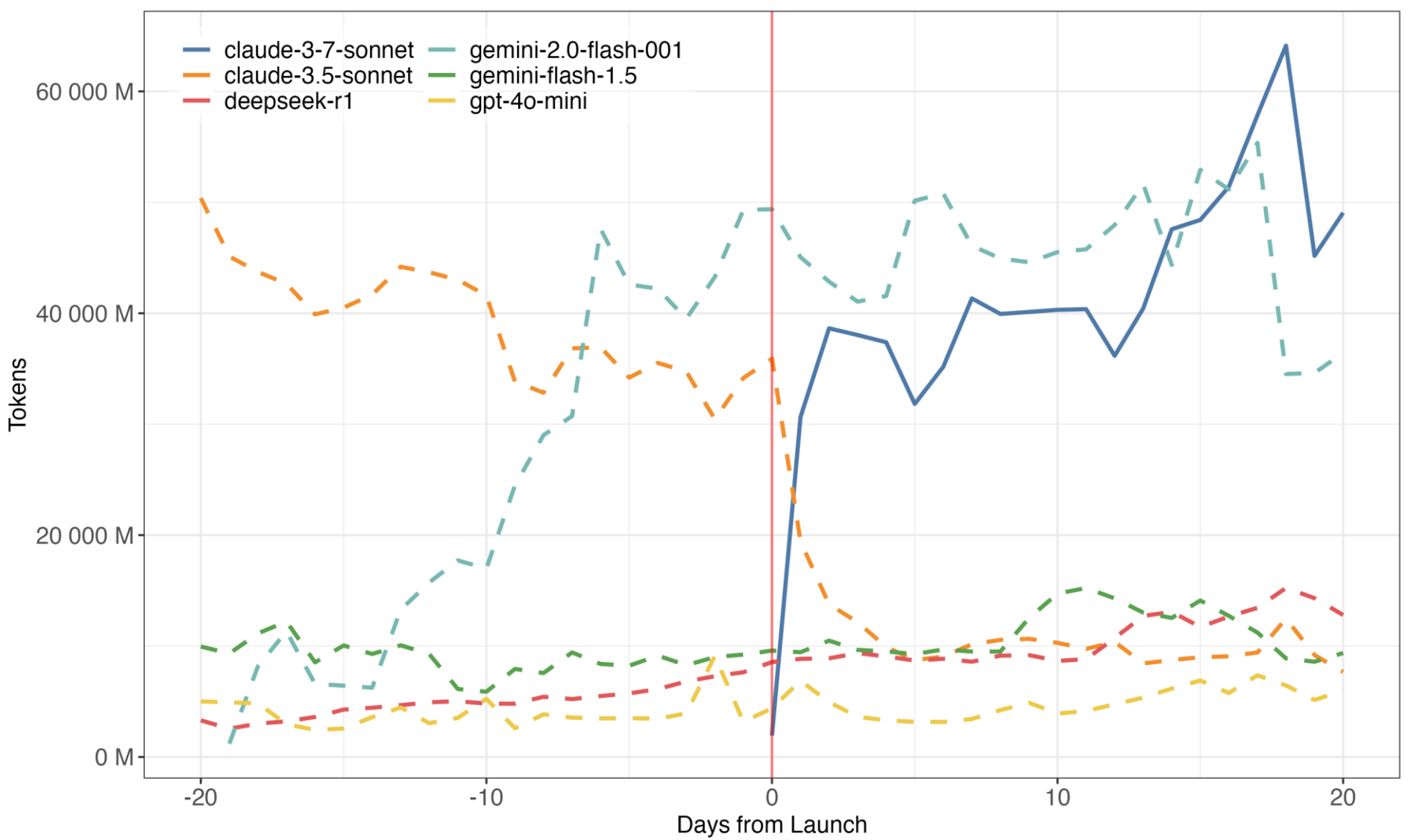


Most multi-homing firms concentrate on one model

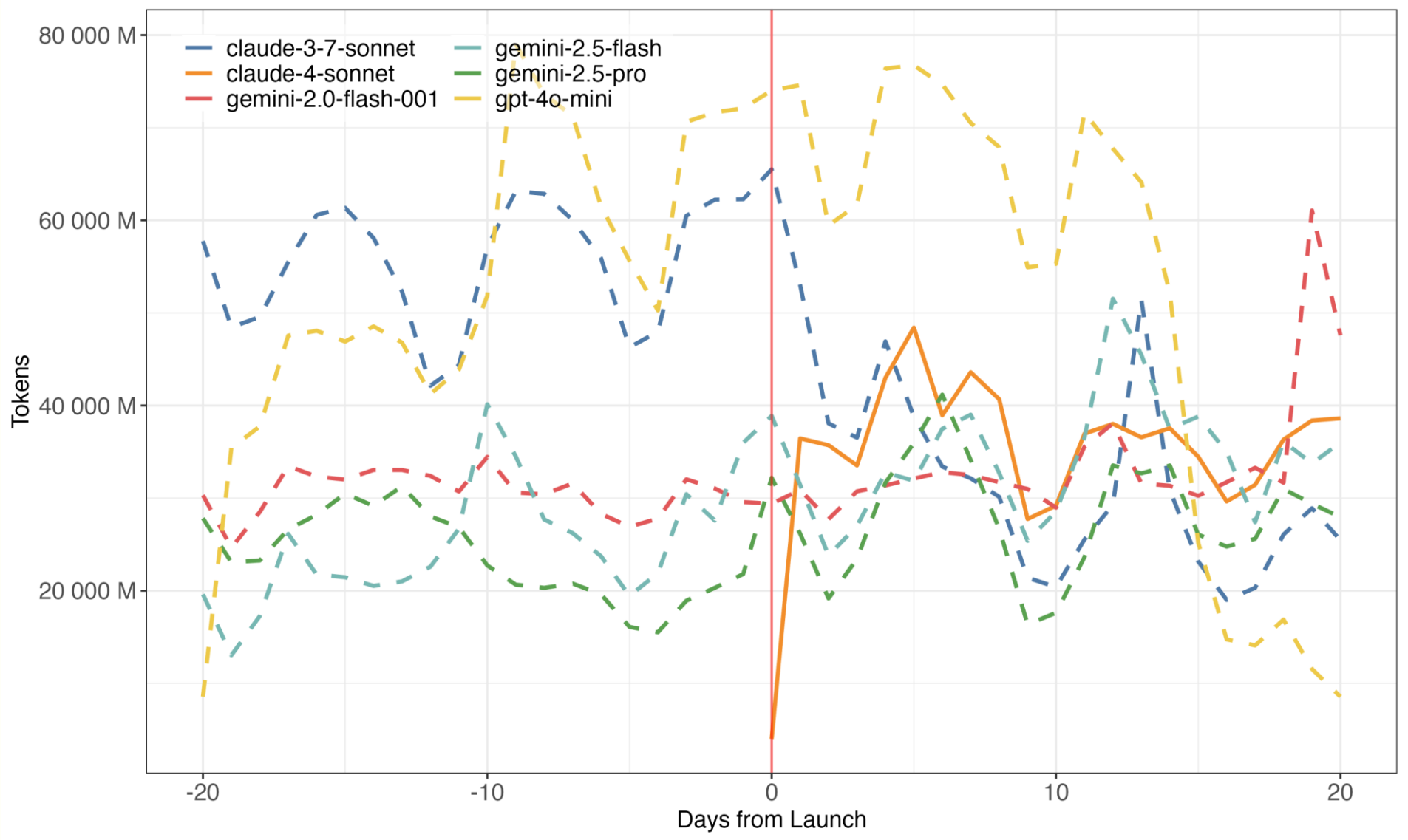
Presentation Roadmap

1. Introduction and Motivation
2. Data Sources and Institutional Background
3. Supply
5. Demand
6. **Substitution and Market Expansion**
7. Usage and Heterogeneity
8. Conclusion

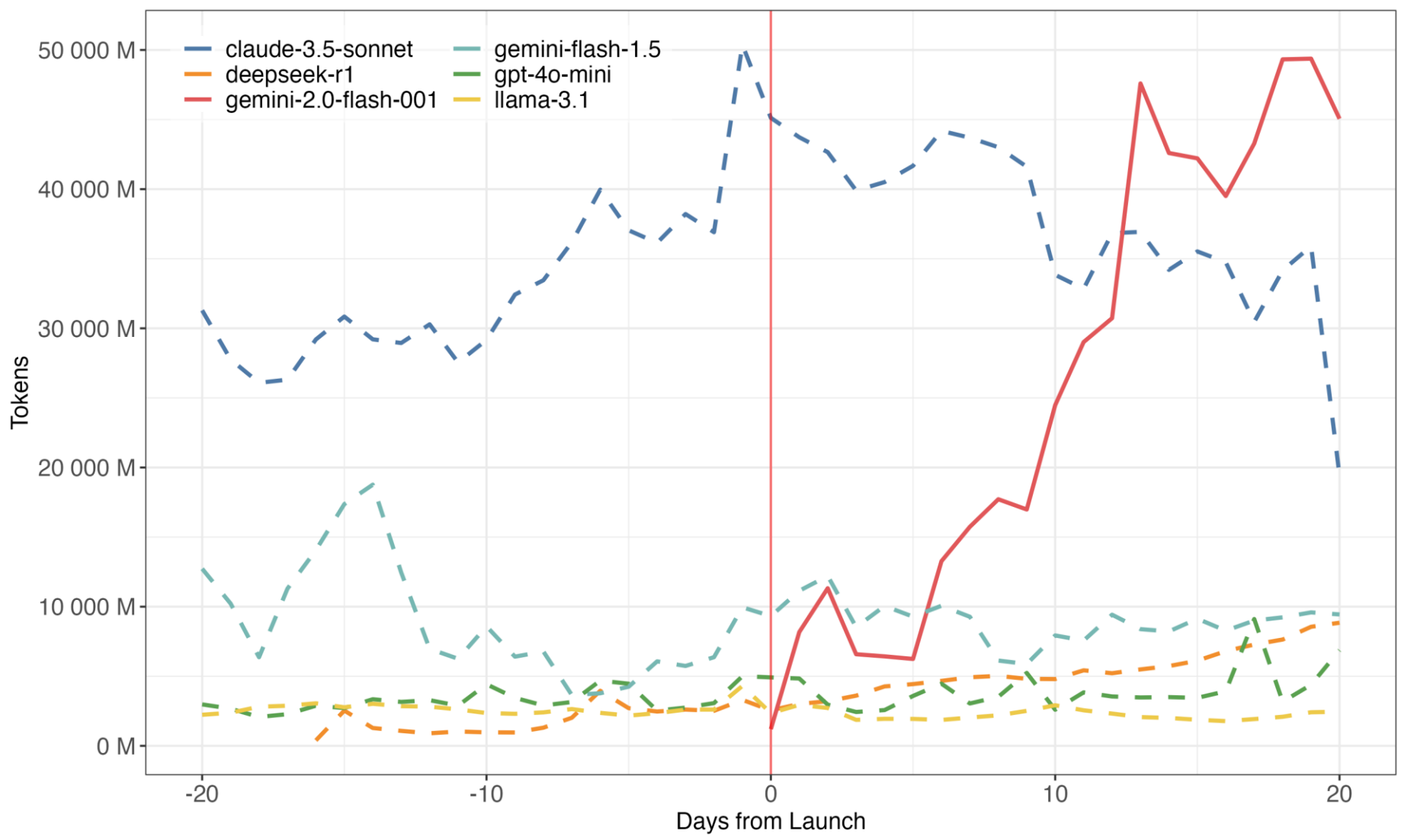
Sonnet 3.7 (Substitution from Sonnet 3.5)



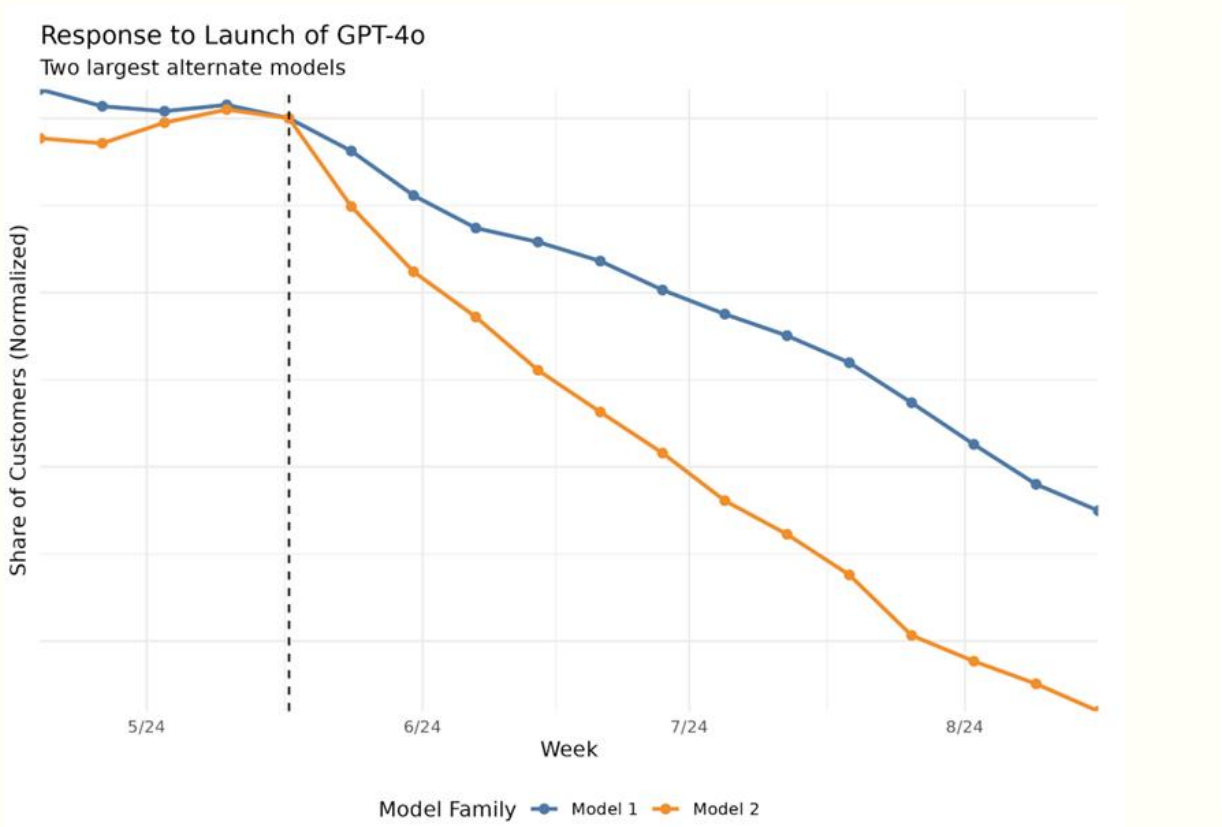
Sonnet 4 (Substitution from Sonnet 3.7)



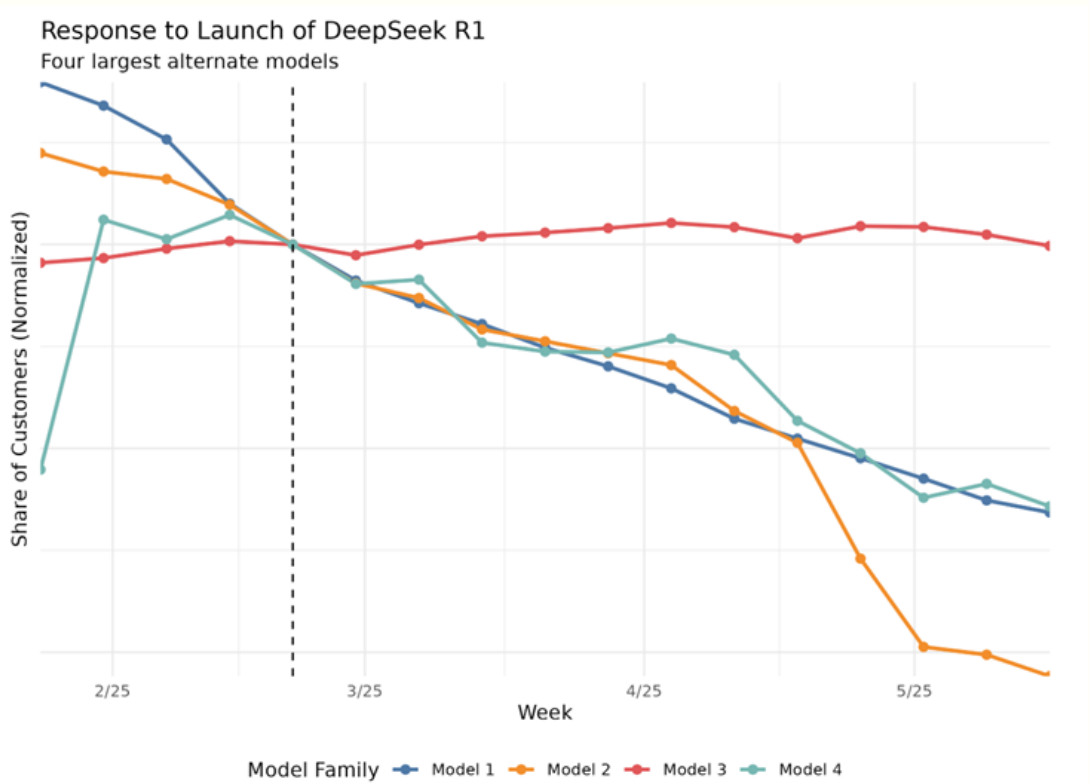
Gemini 2.0 Flash (Market Expansion)



Azure Model Release Event Studies



Gpt-4o Release



DeepSeek R1 Release

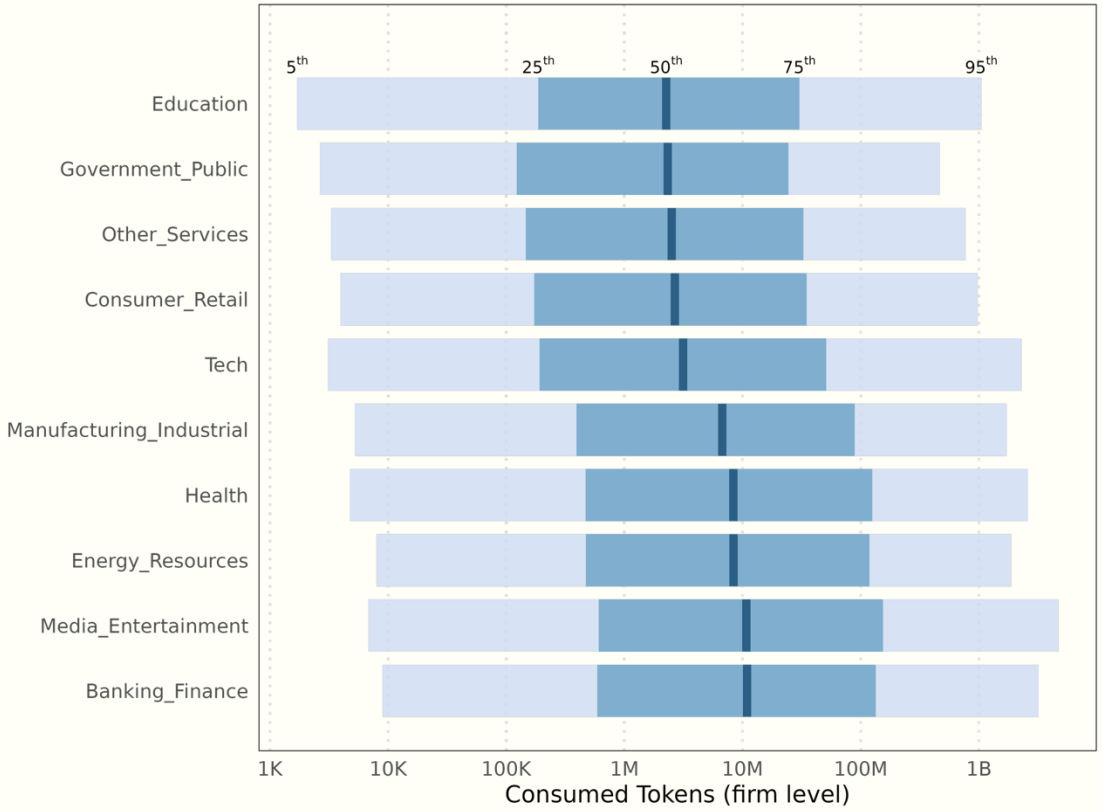
Presentation Roadmap

1. Introduction and Motivation
2. Data Sources and Institutional Background
3. Supply
5. Demand
6. Substitution and Market Expansion
7. **Usage and Heterogeneity**
8. Conclusion

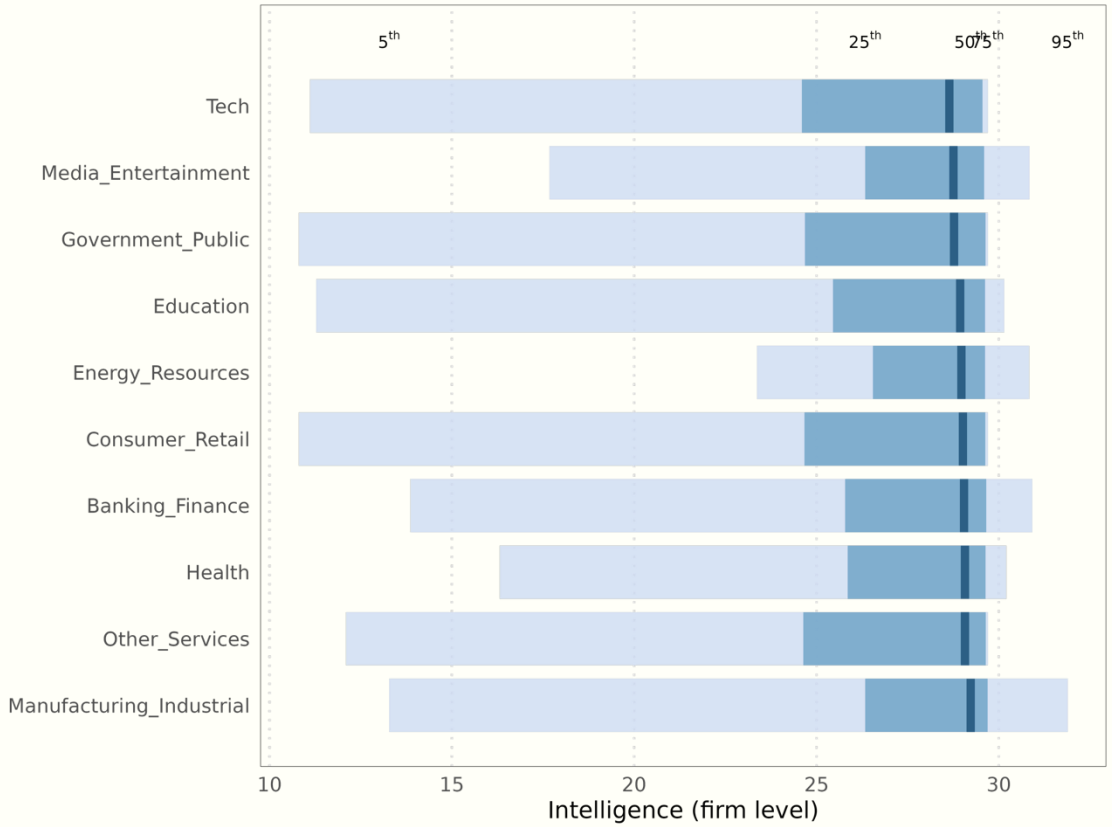
Heterogeneity in diffusion

- Diffusion is critical for understanding and predicting the impacts of AI.
- Diffusion happens on extensive and intensive margins.
- Do firms adopt a new model and how much do they adopt?
- Azure has many enterprises, representing a variety of industries.
- OpenRouter serves many use cases.

Heterogeneity across industries: Azure

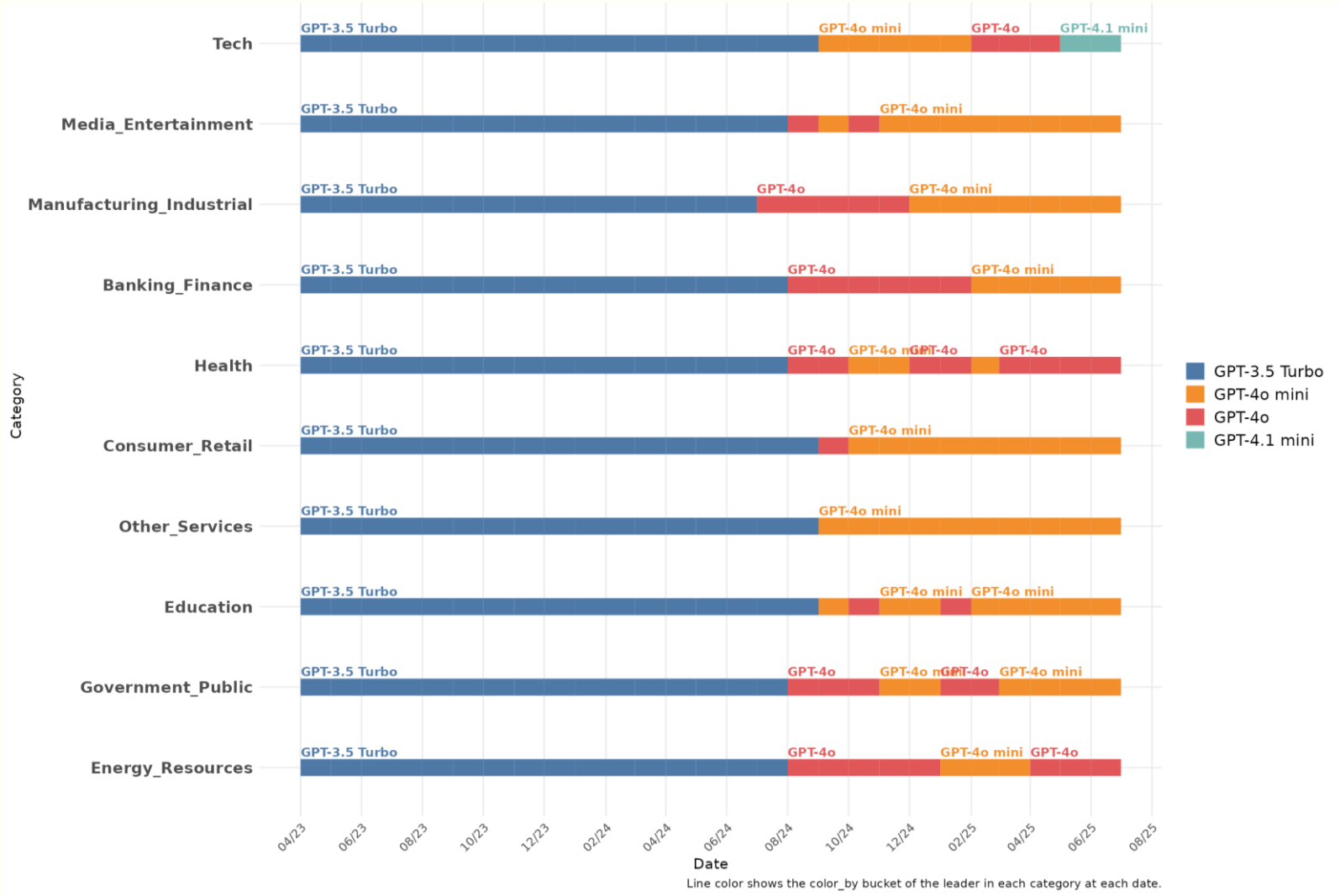


Token Consumption

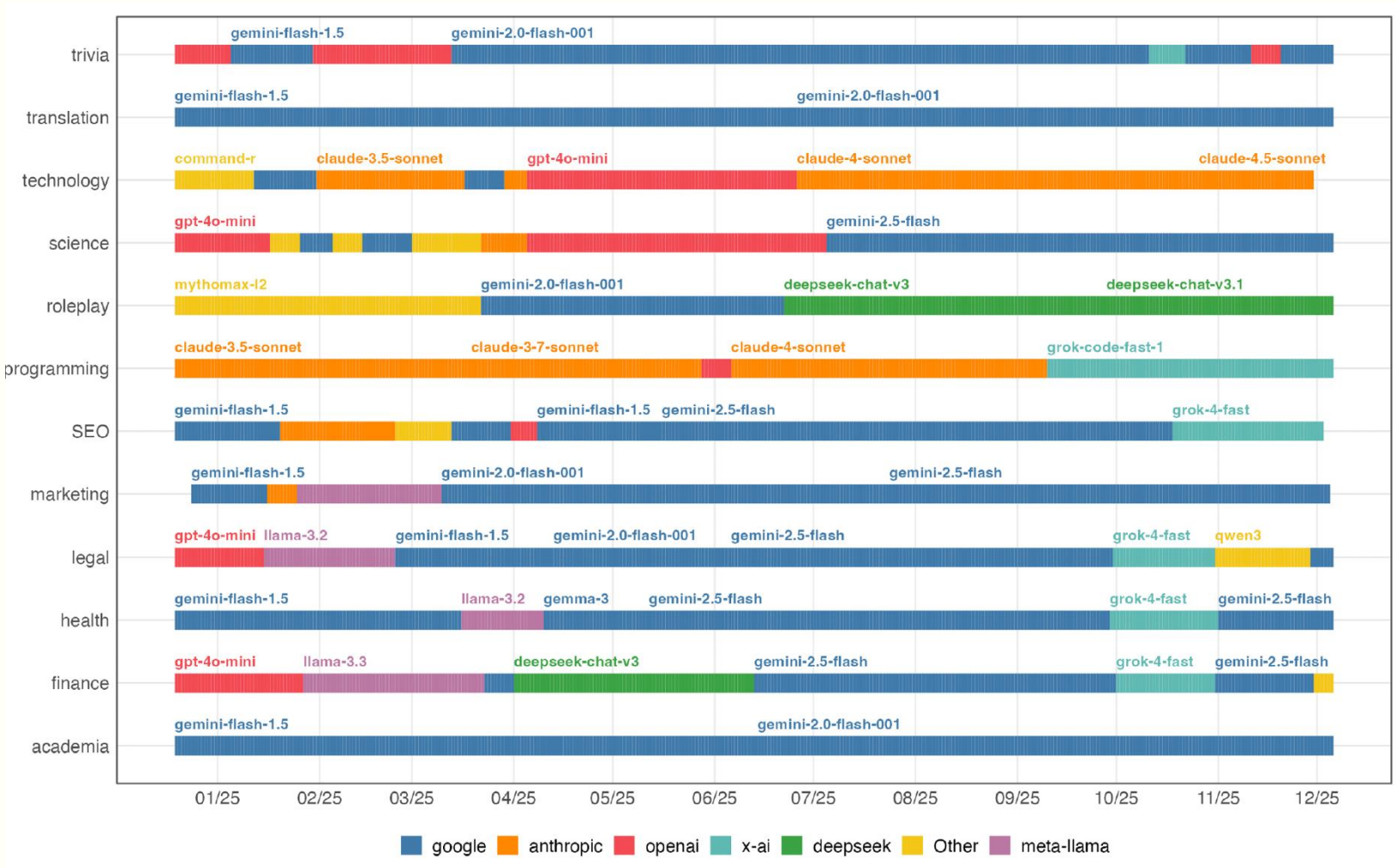


Intelligence

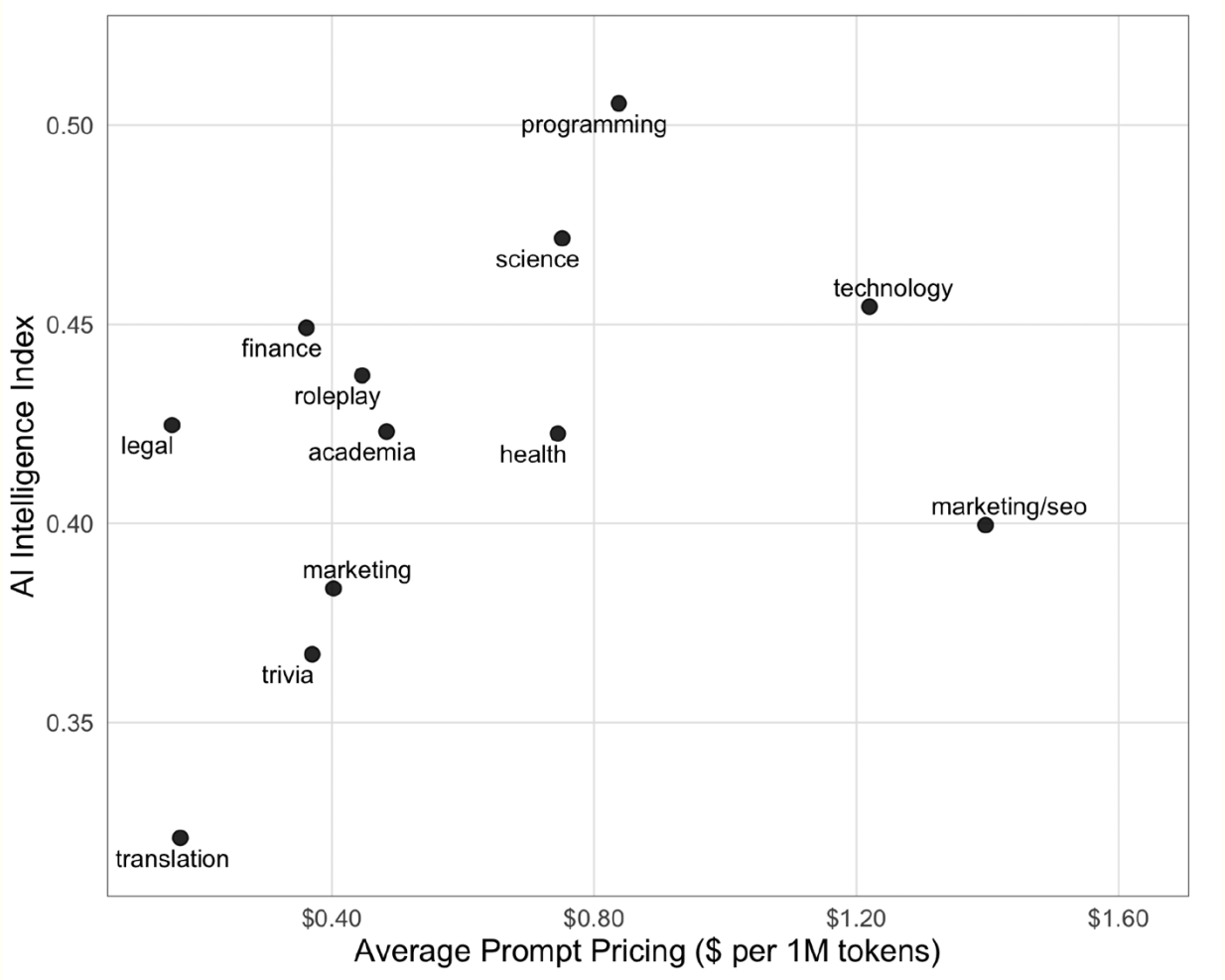
Azure market leadership by vertical



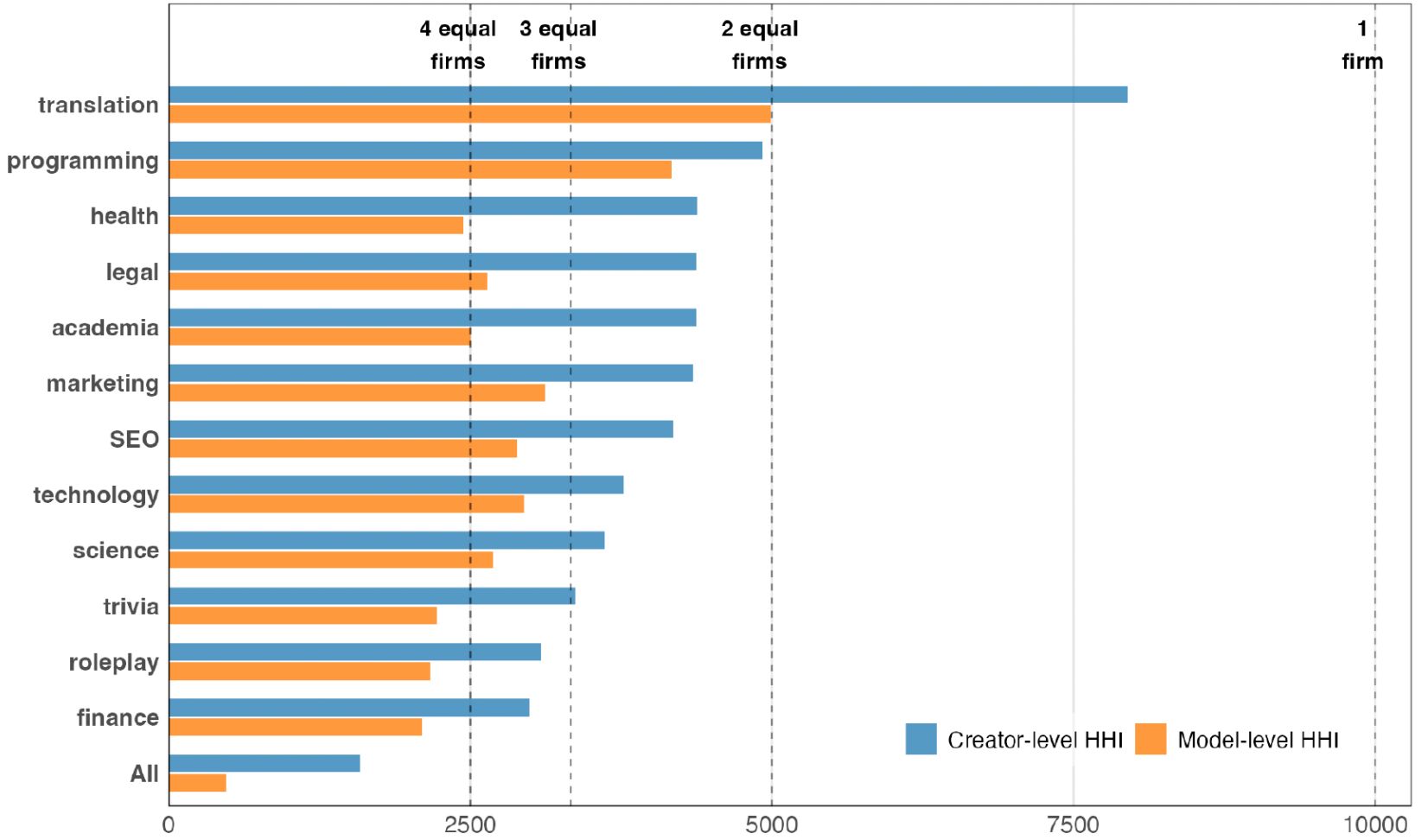
OpenRouter market leaders by use case



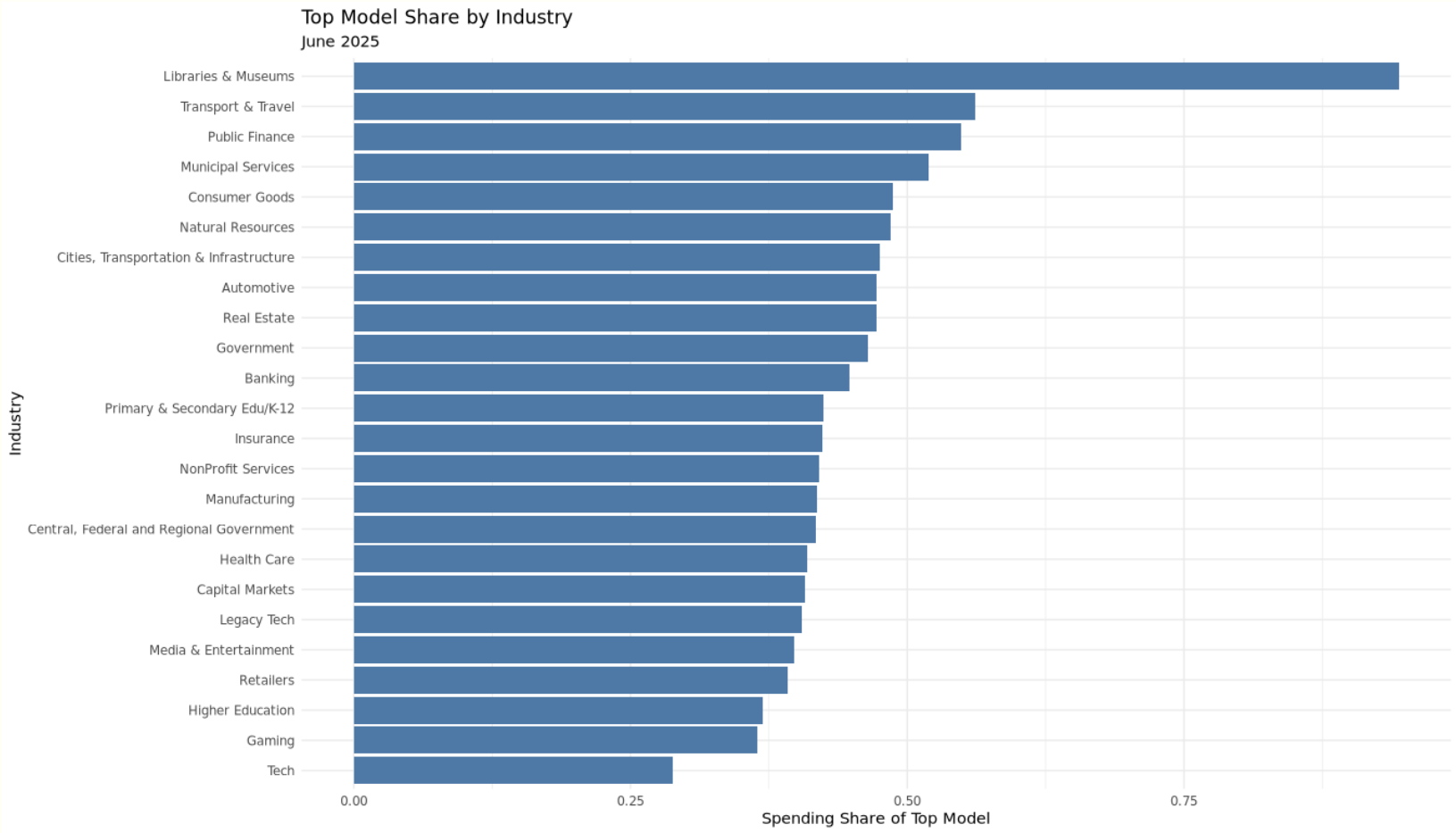
OpenRouter, intelligence and price vary by use case



OpenRouter, concentration indices



Azure, top model share by industry



Summary

- Models are getting better and cheaper at unprecedented rates.
- Closed-source models still have high demand even though they are more expensive for measured intelligence.
- Limited Jevons effects so far.
- Substitution between models is fast, but patterns of substitution suggest model differentiation.
- Concentration indices and switching behavior suggest a competitive industry.

Parting Thoughts

- The market for intelligence may eventually be the biggest market in the world.
- Room for many papers on this and related topics in the IO of AI.
- We have much to contribute on AI topics, but risk irrelevance if we only talk to each other and value top 5's.

Thank You!

afradkin@gmail.com

@AndreyFradkin