AI Premium

Nicola Borri, Yukun Liu, and Aleh Tsyvinski*

July 2, 2026

Abstract

Using 380 trillion tokens of realized AI consumption across more than four hundred large language models from the licensed proprietary OpenRouter dataset covering approximately 2 percent of current global monthly AI token consumption, we analyze how AI affects firms, markets, and workers. Leveraging the unprecedented size, scope and granularity data, we construct the *AI Factor* from growth in tokens, dollars, and users, estimate firm-level *AI Betas* from stock return comovement, and characterize the *AI Premium*. First, we build a high-frequency AI factor and decompose it into salient components. Second, we show that firms whose returns covary more positively with the AI factor—high AI beta firms—earn higher subsequent returns, and the AI premium is large and heterogeneous. A value-weighted long-short strategy earns 64.1 basis points per week, and the premium is large for loadings on the intensive, frontier-oriented margin of AI consumption—closed-source models, paying and seasoned users, and long prompts—but not on casual or open-weight use. Third, the premium reaches beyond technology firms into consumer-facing and capital-heavy parts of the economy, but is absent in emerging markets, including China. Fourth, the AI exposure is more positive in nonroutine interactive work and more negative in analytical, scientific, and operations-control skills—an occupation one standard deviation higher in interaction-and-communication content has 0.36-standard-deviation higher market-implied AI premium. Additionally, we provide early evidence of the rise of the agentic economy.

*Nicola Borri is with LUISS University. Yukun Liu is with the University of Rochester, Simon Business School. Aleh Tsyvinski is with Yale University and NBER. Data provided by OpenRouter, Inc. Used under license. The analysis is based only on aggregated and anonymized derivative data constructed by the authors. The original OpenRouter data are not redistributed or disclosed. We are grateful to OpenRouter for providing access to the data, especially to Justin Summerville for his assistance throughout the data-access process. and to Ron Borzekowski and the Yale Data-Intensive Social Science Center for support. We thank Daron Acemoglu and Andrew Atkeson for detailed comments, Ron Borzekowski and the Yale Data-Intensive Social Science Center for assistance in onboarding the data.

1 Introduction

Using 380 trillion tokens of realized AI consumption across more than four hundred large language models from the licensed proprietary OpenRouter dataset covering approximately 2 percent of current global monthly AI token consumption[1], we analyze how AI affects firms, markets, and workers. Leveraging the unprecedented size, scope and granularity of OpenRouter AI data, we construct and analyze a forward-looking measure of firms' exposure to AI demand, use this measure to identify firms that equity markets consider more exposed to AI, and characterize the market-implied impact of AI on the value of firms, tasks, and skills.

From the universe of OpenRouter data, we construct the *AI Factor*, estimate *AI Betas*, and characterize the *AI Premium* in the cross-section of firms' market values and workers' skills. Specifically, we first construct the *AI Factor*—a high-frequency measure of growth in realized worldwide AI consumption, of which we further analyze its various salient components. We then measure how strongly each public firm's equity price co-moves with this factor—firms' *AI Betas* (or AI exposure)—and characterize how the equity market values that sensitivity, the *AI Premium*.

The paper has four main parts. First, we build a high-frequency, time-varying AI factor from the full universe of OpenRouter AI token consumption data and decompose it into its salient components: frontier closed-source versus open-weight models, core versus new users, seasoned versus casual users, long versus short prompts, prompt-content categories (e.g., programming, science, roleplay), and agentic versus ordinary requests. The AI factor and its salient components show an uneven diffusion, concentration of token consumption in programming, technology, and sciences, and the dramatic increase in agentic share of tokens from negligible to more than half of total AI consumption. Second, we show that AI exposure is priced in the cross-section of stock returns, and the AI premium is both large and heterogeneous. Firms with higher AI betas earn higher average returns. A value-weighted long-short strategy earns 64.1 basis points per week, and the AI premium is on the intensive, frontier-oriented margin of AI consumption—closed-source models, paid/core and seasoned users, and long prompts, rather than on open-weight or casual use. Third, the AI premium reaches beyond technology firms, into consumer-facing and capital-heavy parts of the economy: industry exposure is most positive in retail and consumer durables and turns most negative in health and non-durable goods. We also show that the AI premium is larger in regions where AI development and adoption is more advanced—about 18 basis points per week across developed markets, such as Europe, but absent in emerging markets, including China. Fourth, we build a market-implied mapping of AI exposure onto the tasks and skills across the US economy. We show that the nonroutine, interactive work—communication, persuasion, instruction, and the installation and repair of new systems—loads more on the AI factor, and analytical, scientific, and operations-control skills load less. The most positive loading is on interaction and communication skills: an occupation which is one-standard-deviation higher in interaction-and-communication content has a 0.36 standard-deviation higher market-implied AI exposure.

[1] OpenRouter's May 2026 run rate is about 100 trillion tokens per month (Das and Murphy, 2026). For comparison, Google alone disclosed processing 3.2 quadrillion tokens per month at I/O 2026 (Pichai, 2026), so accounting for the other major providers puts total global LLM token consumption at roughly 5 to 7 quadrillion per month and OpenRouter at about 2 percent, also consistent with the projections by Goldman Sachs Research (Tan, 2026).

Unlike task-based (Acemoglu and Restrepo, 2019b) and usage-based approaches, such as the Anthropic Economic Index (Anthropic, 2026), our measures are constructed from market-implied risk, building on the literature of using stock-market prices to value innovation (Kogan, Papanikolaou, Seru and Stoffman, 2017).

We now provide a detailed description of the underlying data. OpenRouter, Inc. is a global AI platform that routes requests between users and more than four hundred large language models through a single interface—from frontier closed-source models such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini to open-weight models such as Meta's Llama, DeepSeek, and Qwen, to agentic models such as Nous Research's Hermes and coding-specialized models such as Alibaba's Qwen Coder. Our licensed data is built from the universe of OpenRouter data, anonymized and aggregated to the user–model–day level, from January 2024 through April 2026. For each anonymized user, model, and day, we observe the number of requests, the completion and prompt tokens, and the dollar cost, as well as many other characteristics. The panel covers millions of anonymized accounts, allowing us to track individual token consumption of each anonymized user over more than two years. Two additional features of the data are important to note. First, token consumption is realized rather than reported, since every observation is a paid request carrying its own token count and dollar cost. Second, the dataset spans every major provider rather than just a single model: the average user account draws on more than five distinct AI models, and seven in ten use at least two providers. The unprecedented granularity of the data allows us to separate frontier closed-source from open-weight consumption, intensive from casual users, to identify the content category of each request, to determine agentic requests that call external tools, and track a large variety of other salient characteristics of anonymized user token consumption over time and across models. We start by building three weekly series—total tokens, dollars spent, and distinct active users—that measure the quantity, expenditure, and reach of global AI consumption.[2]

First, we build the AI factor using realized AI consumption series: the first principal component of the weekly growth rates of tokens, dollars, and active users. The weekly AI factor loads positively and evenly on all three measures. This aggregate index is our baseline factor, but the user–model–day granularity also allows us to construct separate sub-factors for salient components of token consumption by model tier, user intensity and tenure, prompt complexity, content category, and agentic use, to estimate which components of AI consumption are priced by the equity markets.

Second, we show that AI exposure is priced in the cross-section, with an AI premium that is large and heterogeneous. Firms whose returns covary more positively with the AI factor—high AI beta firms—earn higher subsequent returns: in value-weighted quintile portfolios the most exposed firms outperform the least exposed by 64.1 basis points per week ($t = 2.84$), about 56 basis points after Fama-French five-factor and momentum adjustment ($t = 2.41$). The Fama-MacBeth regressions also show a positive price of AI risk after controlling for size, value, profitability, investment, momentum, reversal, leverage, and accruals. Two caveats qualify this interpretation. First, the sample covers an early, fast-moving phase of AI diffusion, so the estimates should be interpreted as the current price of AI exposure, rather than a long-run estimate of the AI premium. Second, data from OpenRouter, Inc., though exceptionally rich, likely overrepresents

[2]For comparison, the Anthropic Economic Index, currently the most prominent and comprehensive public measure of AI demand, classifies only about one million Claude conversations from a single provider over a short sample period (Anthropic, 2026).

sophisticated and developer users, so the premium is tied to the market pricing of realized observed AI consumption influenced toward these users.

The sign of the estimated AI premium is consistent with a transition-risk channel of technological change (Pastor and Veronesi, 2009). New technologies reallocate rents across firms, sectors, capital vintages, and workers. Innovation can destroy the value of older technologies even as it creates new ones (Aghion and Howitt, 1992), and when risky growth opportunities are imperfectly diversified, risk-averse investors require higher expected returns to hold them (Acemoglu and Zilibotti, 1997). Our result that the AI premium is positive and quantitatively sizable is most consistent with the argument that when this reallocation is systematic, investors require compensation for holding the firms most exposed to it (Pastor and Veronesi, 2009). Important papers by Acemoglu and Restrepo (2019b, 2022) formalize the task replacement mechanism: automation displaces labor from some tasks, the creation of new tasks can reinstate labor elsewhere and redistribute gains across workers and firms. Our results on the market-implied betas of occupations, skills, and tasks provide a different measure of labor exposure from the point of view of equity markets rather than from the view of technological possibility of automation.

We then test robustness to a possible alternative that the AI factor merely repackages the recent rally or potential hype in technology stocks and AI-themed equities. The evidence does not support this view. When the first-stage exposure regression controls for high-technology stocks, a value-weighted semiconductor portfolio, or an AI/robotics ETF basket, the value-weighted high-minus-low spread remains 49.5 basis points ($t = 2.49$), 45.4 basis points ($t = 2.23$), and 63.4 basis points ($t = 3.11$) per week, respectively. The spread is also 60.3 basis points ($t = 3.03$) after demeaning exposures by Fama-French 30 industry and 69.6 basis points ($t = 3.35$) after controlling for Google Trends attention to AI. These tests show that the AI premium prices exposure to realized AI consumption, and not exposure to technology stocks or AI-themed attention.

Third, we turn from the existence of the AI premium to where the AI premium is stronger. We start by splitting AI token consumption into its salient components. The AI premium loads on the intensive margin of AI consumption—frontier models, experienced users, and complex tasks—rather than on casual or extensive use. The AI premium is larger when exposure is estimated from frontier closed-source model consumption: the high-minus-low spread is 53.4 basis points per week ($t = 2.47$), compared with 32.3 basis points ($t = 1.76$) for open-weight models. Similar results also hold across the other measures of sophisticated AI consumption: 66.8 basis points ($t = 2.37$) for paid/core users versus 21.9 ($t = 0.86$) for new users, 59.3 ($t = 2.13$) for seasoned users versus 29.3 ($t = 1.21$) for non-seasoned users, and 54.1 ($t = 2.40$) for long prompts versus 33.0 ($t = 1.79$) for short prompts. Next, we examine the AI premium by countries. Repeating the portfolio sorting exercise gives a 17.9-basis-point weekly spread ($t = 2.77$) in developed markets, but only 5.0 basis points ($t = 0.94$) in emerging markets, including China. These results shows that the premium is more pronounced in regions where listed firms and investors are closer to frontier AI development and adoption.

Furthermore, we present two complementary results to investigate whether part of the observed average excess return might reflect a series of AI surprises rather than a priced risk, which may be important for a technology as new as AI. In our event study of AI model releases, high-AI-beta firms outperform low-AI-beta firms around frontier-provider releases by 1.9 percent over the five-day window ($t = 4.17$), and by

1.1 percent after five-factor and momentum adjustment ($t = 2.20$), with muted responses to non-frontier releases. When we instead remove the frontier model-release weeks from the sample, the value-weighted high-minus-low spread remains positive at 0.395 percent per week. The premium is stronger around releases but it accrues gradually throughout the sample rather than only on news dates.

Fourth, we use the same baseline firm-level exposures to determine market-priced AI exposure to occupations and skills in the current organization of work. We first map firm and industry exposures into occupations using BLS employment weights and then map occupations into skills using O*NET ratings. Then, we compare the AI exposures of occupations with the task taxonomies in Autor, Levy and Murnane (2003) and Acemoglu and Autor (2011), and the skill measures in Deming (2017). We thus construct a market-priced map of AI exposure for the occupations, which is different from the technical automability score or labor-displacement forecast that are the focus of task- and capability-based exposure measures such as (Eisfeldt, Schubert and Zhang, 2023; Kogan, Papanikolaou, Schmidt and Seegmiller, 2023). The AI exposure is more positive in installation and repair, programming, persuasion, instruction, and systems-integration occupations, and it is more negative in science, healthcare, and operations-control occupations. In terms of tasks, a one-standard-deviation increase in nonroutine interactive content is associated with 0.15 standard deviations higher market-implied AI exposure ($t = 2.87$), while nonroutine analytic content is associated with 0.15 standard deviations lower exposure ($t = -3.41$). In terms of skills, social skills enter positively at 0.16 ($t = 3.44$), and the strongest detailed loading is interaction and communication, 0.36 ($t = 4.21$); information use enters negatively at -0.29 ($t = -2.43$). Existing occupation-level AI-exposure measures explain less than 2 percent of this variation.

Finally, we provide early evidence on the rise and market-implied effect of agentic AI consumption, defined as models executing multi-step workflows that plan, call external tools, take actions, and coordinate with other agents (see, e.g., Hassabis and Kavukcuoglu, 2024; Ng, 2024a,b; Yang, Yonack, Zyskowski, Yarats, Ho and Ma, 2025). We measure this component in the OpenRouter data as requests in which the model invokes an external tool. We document the rise in the agentic token share from near zero in 2024 to roughly half of all tokens by the end of the sample, and the agentic dollar share rises more slowly as the realized price per agentic token falls. This decline in unit cost is consistent with more efficient routing of agentic requests to cost-effective backing models as firms struggle with skyrocketing AI token budgets. The granularity of the OpenRouter data allows us to track three salient token-level components of this activity: tool calls, the model's internal reasoning tokens, and cache reads of recurring prompt prefixes. We show that tool-call and cache-read shares rise to roughly two-fifths to one-half of total tokens, while reasoning tokens rise more slowly. In the cross-section of equity returns, exposure to these agentic-consumption components has positive and economically sizable point estimates of about 0.3 to 0.5 percent per week for the individual agentic factors. Because estimates are imprecise since agentic token consumption consumption is a very recent phenomenon, we interpret this as early evidence of a positive agentic premium.

We next briefly relate our paper to the relevant literature. Acemoglu and Restrepo (2019a,b, 2022) provide the task-based mechanism through which automation displaces labor from some tasks, creates new tasks, and reallocates rents across workers and firms. In asset pricing: Kogan and Papanikolaou (2014) link technology shocks and growth opportunities to expected returns, Garleanu, Kogan and Panageas (2012)

show that displacement risk can be priced, and Pastor and Veronesi (2009) study how technological revolutions are capitalized in stock prices. Recent work also studies AI and market beliefs in asset prices (Andrews and Farboodi, 2025). Directly on AI risk, Babina, Fedyk, He and Hodson (2024a) show that firms investing more in AI experience higher market betas, consistent with AI creating growth options. Babina, Fedyk, He and Hodson (2024b); Babina (2026) measure firms' AI efforts and study their effects on growth, innovation, and broader economic outcomes. An important paper by Eisfeldt, Schubert and Zhang (2023) uses AI to classify task exposure and then maps those task scores to occupations and firms. They show that high-exposure firms rose by about 5 percent relative to low-exposure firms after OpenAI introduced ChatGPT, initially based on GPT-3.5; that the effect is distinct from product-market AI exposure and stronger for firms with data assets; and that labor demand, wages, job postings, and profitability point to labor substitution through exposed core tasks. Our event-study evidence shows similar results. Demirer, Fradkin, Tadelis and Peng (2025) use early scraped, aggregate data from the OpenRouter's public website to study LLM pricing and demand. The OpenRouter–a16z report by Aubakirova, Atallah, Clark, Summerville and Midha (2025) documents stylized facts about the structure of the token consumption market from a 100-trillion-token OpenRouter data: substantial open-weight adoption, heavy use for roleplay and coding, the rise of agentic inference, global variation, cost-use patterns, and persistent early-user cohorts.

2 Data

This section describes the data used in the empirical analysis. Section 2.1 introduces the OpenRouter data. Section 2.2 describes the remaining data used in the empirical analysis.

2.1 OpenRouter Data

We measure AI consumption (which we refer to interchangeably as *AI use*) using a panel of realized inference requests from OpenRouter, Inc., an AI inference provider that connects users and developers to hundreds of large language models through a single API endpoint. Users submit a prompt and select which model runs it—an OpenAI, Anthropic, Google, Meta, DeepSeek, or open-weight model among more than four hundred available—and OpenRouter routes the request to a backing provider that hosts that model. One benefit of the data is the large number of models available on OpenRouter making the data particularly useful for measuring aggregate AI consumption: observed consumption is unlikely to reflect the product cycle of any single provider, the design of any single consumer application, or the marketing strategy of any single model lab (Aubakirova, Atallah, Clark, Summerville and Midha, 2025). The data are also realized rather than reported AI consumption—every observation is an actual paid request.

We work with user-level OpenRouter data structured as a panel at the user–model–day level, where model identifies a specific model version (e.g., Anthropic's Opus 4.7 is a separate unit from the earlier Opus 4.6). Each observation summarizes the activity of a single anonymized user with a particular model on a particular day: for example, the total requests, prompt and completion tokens, and dollar cost generated by a particular user when interacting with Opus 4.7 on a given day. Importantly, the panel structure allows us to track each anonymized user's consumption over time, across models, and across providers. Starting

from this granularity, we construct three weekly aggregate series that capture salient components of AI consumption: total tokens (the quantity of language-model work demanded), dollar usage (the expenditure on that work, combining quantity and the actual price paid), and distinct users (the extensive margin of adoption). Our sample covers the weekly period from January 2024 through April 2026. Over this time period, OpenRouter intermediates 380 trillion cumulative tokens, which, to our knowledge, is the largest cross-provider record of realized AI consumption available outside the model labs themselves.

The user–model–day granularity is one feature that distinguishes our panel from alternative existing datasets, which are typically tied to a single model, provider, or product. The resulting panel is both large and granular, with the dataset covering millions of anonymized accounts. Most accounts also use multiple models and providers: the average account draws on more than five distinct models, and seven in ten accounts use at least two providers (see Online Appendix Table IA.1). The granularity of the data allows us to decompose AI consumption along multiple margins we exploit in the asset-pricing tests: frontier closed-source models versus open-weight models, paid/core users versus the rest, seasoned users versus new arrivals, and long, complex prompts versus short ones. The same granularity supports classification of activity by prompt-content category and identification of requests that invoke external tools (e.g., agentic use).

Our baseline AI factor aggregates the three main AI distinct consumption series into a single index, defined as the first principal component of their standardized weekly log growth. We use growth rather than levels because asset returns should respond to the unexpected component of AI consumption—changes in adoption and intensity—rather than to the mechanical upward trend of a diffusing technology. The first principal component places weights of 0.665, 0.559, and 0.496 on the three series and accounts for 56.5 percent of their joint variance, so a single common factor captures the bulk of the comovement across the three margins (see Online Appendix Table IA.2). Section IA.3 in the Online Appendix also reports asset pricing results based on the simple cross-sectional average, and each component series separately as robustness checks, with similar conclusions.

Figure 1 plots weekly total-token consumption on a log scale. Total tokens rise from 11.4 billion in the first week to 15.6 trillion in the final observed week, with cumulative total-token consumption of 380.8 trillion over the indexed sample. The growth persists during the sample and is not concentrated around a single model release, which is consistent with steady diffusion of the technology across users and applications.

Table 1 reports full-sample summary statistics for the AI variables (Panel A) and the firm characteristics used in the asset-pricing tests (Panel B). Panel A reports the baseline factor and its three standardized input growth series, along with the AI sub-factors computed on different subsets of the data: AI^{Closed} and AI^{Open} from closed-source and open-weight model consumption, AI^{Core} and AI^{New} from paid/core and new accounts, $AI^{Seasoned}$ and $AI^{Nonseasoned}$ from seasoned and non-seasoned users, and $\Delta \ln Tok^{Long}$ and $\Delta \ln Tok^{Short}$ from long- and short-prompt total-token growth. We use these additional factors in the analysis of Section 3.

Section IA.6 in the Online Appendix reports additional descriptive information on other aspects of the data. Figure 5 decomposes total tokens into agentic and non-agentic consumption. Figure IA.1 reports the share of weekly tokens by prompt-content category for all models and separately for closed- and open-

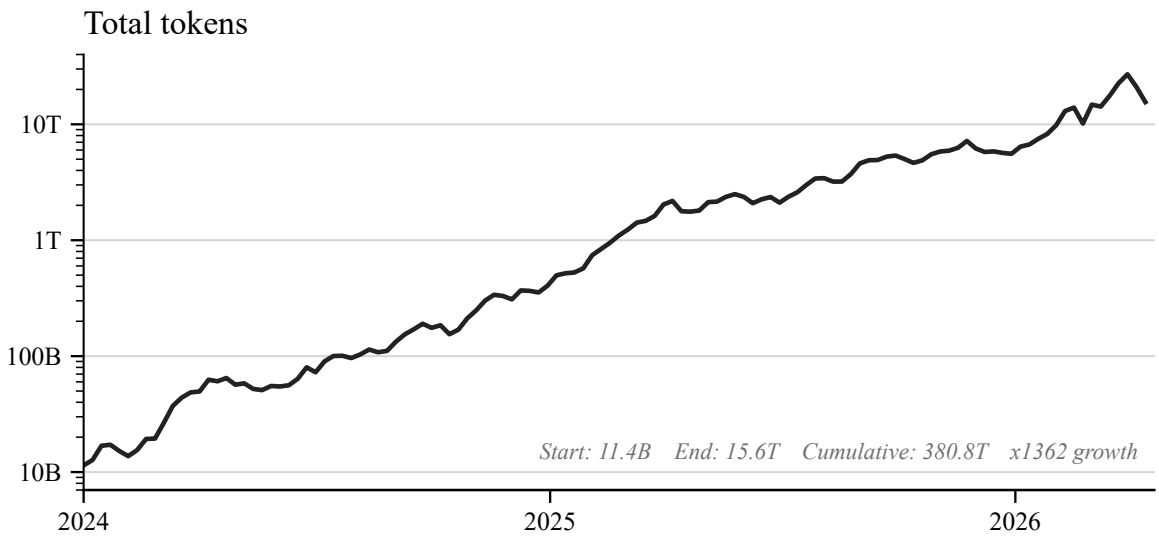


Figure 1: Weekly Total Tokens

The figure plots weekly total-token consumption on a log scale. Total tokens are prompt tokens plus completion tokens. The in-figure annotation reports the first-week level, final observed-week level, cumulative total-token consumption of 380.8 trillion, and growth multiple. The corresponding standardized weekly log growth series is plotted in Online Appendix Figure IA.2.

Table 1: Summary Statistics

	Mean	Std.	p10	p25	p50	p75	p90
Panel A. AI variables							
AI	0.000	1.302	-1.561	-0.872	-0.107	0.726	1.596
$\Delta \ln Tok$	0.000	1.000	-1.274	-0.555	0.050	0.608	1.249
$\Delta \ln Dol$	0.000	1.000	-1.299	-0.609	-0.052	0.581	1.233
$\Delta \ln User$	0.000	1.000	-1.168	-0.607	0.004	0.525	1.150
AI^{Closed}	0.000	1.341	-1.555	-0.911	-0.177	0.816	1.381
AI^{Open}	0.000	1.297	-1.422	-0.784	-0.103	0.720	1.642
AI^{Core}	0.000	1.272	-1.343	-0.849	-0.028	0.773	1.362
AI^{New}	0.000	1.271	-1.516	-0.863	0.153	0.898	1.537
$AI^{Seasoned}$	0.000	1.276	-1.323	-0.829	-0.061	0.818	1.312
$AI^{Nonseasoned}$	0.000	1.254	-1.748	-0.703	-0.034	0.704	1.701
$\Delta \ln Tok^{Long}$	0.069	0.183	-0.132	-0.053	0.064	0.184	0.295
$\Delta \ln Tok^{Short}$	0.053	0.142	-0.079	-0.030	0.039	0.137	0.225
Panel B. Firm characteristics							
Ret_{t+1}	0.231	9.211	-6.504	-2.683	0.100	2.902	6.674
$\ln ME$	14.41	2.06	11.74	12.92	14.39	15.79	17.20
$\ln BM$	-0.90	1.00	-2.24	-1.50	-0.75	-0.14	0.22
$Profit$	0.239	0.253	0.018	0.055	0.206	0.363	0.551
Inv	0.092	0.324	-0.109	-0.025	0.035	0.117	0.285
Mom	0.081	0.415	-0.387	-0.124	0.080	0.291	0.534

Panel A reports full-sample summary statistics for the weekly AI variables, and Panel B reports full-sample summary statistics for the firm characteristics. AI is computed as the first principal component of standardized weekly log growth in total tokens ($\Delta \ln Tok$), usage dollars ($\Delta \ln Dol$), and distinct users ($\Delta \ln User$). The table further reports summary statistics for AI variables based on closed- and open-source models, paid/core and new users, seasoned and non-seasoned users, and long- and short-prompt total-token growth.

source models. Figure IA.3 plots the token-share evolution of the four model- and user-side decompositions used to compute the additional factors reported in Panel A of Table 1.

2.2 Additional Data

For the U.S. equity analysis, we use CRSP common stocks listed on the NYSE, AMEX, and Nasdaq matched to Compustat accounting data. The baseline sample keeps firms with formation-week market capitalization of at least $1 million and CRSP price of at least $5, and we compute weekly returns and the standard firm characteristics: log market equity, log book-to-market, gross profitability, investment, and momentum. For factor adjustments we use the Fama-French five factors (Fama and French, 1993, 2015) and the momentum factor (Jegadeesh and Titman, 1993; Carhart, 1997). For industry adjustments we use the Fama-French 30 industry classification. For technology-sector returns we use the Fama-French 10 high-technology industry return from Kenneth French's data library. We also construct a semiconductor portfolio return as the value-weighted return on CRSP common stocks in three semiconductor-related NAICS industries, as semiconductors are crucial for the development of the AI infrastructure.[3]

For the international analysis, we use Compustat Global common stocks split into MSCI developed and emerging markets, applying the same market-capitalization and price filters in U.S. dollars and measuring returns in local currency. For the China test, we use the Datastream China A-share dataset together with a China-specific AI factor built from mainland-China OpenRouter consumption. To capture broader AI attention and from traded AI exposure, we use Google Trends search interest for the topic "Artificial intelligence" worldwide and an equal-weighted basket of seven AI- and robotics-themed exchange-traded funds.[4] Finally, for the event-study analysis we hand-collect a calendar of major AI model releases from official model cards (the labs' technical release documentation), company blog posts, and developer announcements, separating releases by the five frontier providers—Anthropic, DeepSeek, Google, Meta, and OpenAI—from those of other providers. The full calendar with sources is reported in Online Appendix Table IA.16.

3 The AI Premium and the Cross-Section of Equity Returns

This section examines whether firms' AI exposure is priced in the cross-section of equity returns. Section 3.1 forms portfolios sorted on firm-level AI exposure and documents the AI premium. Section 3.2 presents the results of Fama-MacBeth (Fama and MacBeth, 1973) cross-sectional regressions. Section 3.3 reports results on robustness tests, international evidence, and heterogeneity across subcomponents of AI consumption. Section 3.4 presents firm-level exposures within the S&P 500.

3.1 AI Portfolios

We define firm i's market-implied AI exposure at week t, $\beta_{i,t}^{AI}$, as the coefficient estimate from a rolling weekly regression of its log excess return on the baseline AI factor, controlling for the log excess market

[3]NAICS 334413 (Semiconductor and Related Device Manufacturing), 333242 (Semiconductor Machinery Manufacturing), and 334418 (Printed Circuit Assembly Manufacturing).

[4]The ETF basket includes the following ETFs: BOTZ, AIQ, IRBO, ROBO, ARTY, WTAI, and CHAT.

return:

$$ret_{i,\tau} = \alpha_{i,t} + \beta_{i,t}^{AI} \times AI_\tau + \beta_{i,t}^{rm} \times r_{m_\tau} + \epsilon_{i,\tau}, \quad \tau \in \mathcal{W}_t, \tag{1}$$

where $\mathcal{W}_t$ is the 13-week estimation window ending at formation week t, with a minimum availability requirement of 9 weeks. The t subscript on $\beta_{i,t}^{AI}$ indicates that exposure is re-estimated every week and can change as investors revise forward-looking valuations in response to AI-consumption shocks. Our exposure measure is time-varying and is inferred from realized stock-price comovement. Because each exposure uses only returns through the formation week, the AI exposure measure is free of look-ahead.

Portfolios are formed by sorting all firms at the beginning of each week into quintiles by their current AI exposure $\beta_{i,t}^{AI}$, with breakpoints based on the full all-stock universe, and are rebalanced weekly. Our baseline portfolios are value-weighted, and equal-weighted portfolios and NYSE-breakpoint variants are also reported. NYSE breakpoints follow the standard Fama-French sorting convention and reduce the influence of smaller, less liquid non-NYSE stocks on the portfolio cutoffs (Fama and French, 1992). Each quintile contains about 680 stocks on average.

Table 2 reports the properties of the quintile portfolios. The first row shows a large and statistically significant spread in AI exposure across the portfolios. The high-low spread in $\beta^{AI}\sigma_w(AI)$ is 0.0573, with a t-statistic of 35.43. The five quintiles are otherwise similar in size, book-to-market, profitability, and investment: average $\ln ME$ is 13.98 in the lowest-β^{AI} quintile and 14.28 in the highest, $\ln BM$ is -0.96 versus -0.94, profitability is 0.226 versus 0.253, and investment is 0.121 versus 0.118. The small cross-quintile differences in $\ln ME$ and profitability are statistically significant but economically modest. The one characteristic that differs systematically across the quintiles is past-month momentum: Mom rises from 0.019 in the lowest-β^{AI} quintile to 0.140 in the highest, a 12 percent spread that is highly significant. This pattern motivates the inclusion of momentum among the factor-model controls in the pricing tests below.

Table 3 reports the portfolio sorting results. Panels A and B use value-weighted portfolios, so larger firms receive more weight. The results are very similar across the two breakpoint choices. With all-stock breakpoints in Panel A, the high-low excess-return spread is 64.1 basis points per week, with a t-statistic of 2.84. With NYSE breakpoints in Panel B, the spread is 53.3 basis points per week, with a t-statistic of 2.51. Panels C and D repeat the exercise with equal-weighted portfolios, so each firm receives the same weight. The spreads are smaller but remain positive and statistically significant, at 35.0 and 30.9 basis points per week, hence the result is not driven only by a few mega-cap technology firms.

The main finding of this section is that firms with higher AI exposure earn higher future returns than firms with lower AI exposure. This positive and significant high-low spread is the portfolio evidence of an *AI premium*: investors require compensation for holding firms more exposed to AI. The direction of the premium is consistent with the transition-risk channel described in the introduction (Aghion and Howitt, 1992; Acemoglu and Zilibotti, 1997; Pastor and Veronesi, 2009). In simple terms, according to this view of AI, the technology creates new opportunities, but it also changes which firms, sectors, and tasks gain or lose value. During this transition, investors treat AI exposure as a risk to be compensated for. The spread also is robust to standard risk adjustment. In Panel A, the alpha from the Fama–French five-factor model (Fama and French, 2015) is 56.3 basis points per week, and the alpha from the five-factor model augmented with

Table 2: Properties of U.S. AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
$\beta^{AI}\sigma_w(AI)$	-0.0316	-0.0087	-0.0005	0.0071	0.0257	0.0573
t(H−L)						(35.43)
$\ln ME$	13.98	14.64	14.60	14.72	14.28	0.29
t(H−L)						(5.61)
$\ln BM$	-0.96	-0.91	-0.88	-0.87	-0.94	0.02
t(H−L)						(0.48)
$Profit$	0.226	0.245	0.238	0.236	0.253	0.027
t(H−L)						(3.94)
Inv	0.121	0.082	0.074	0.079	0.118	-0.003
t(H−L)						(-0.47)
Mom	0.019	0.073	0.088	0.110	0.140	0.121
t(H−L)						(6.80)
$\sigma(R)_{13w}$	0.087	0.049	0.042	0.048	0.078	-0.009
t(H−L)						(-5.87)
Analyst disp.	0.373	0.215	0.171	0.178	0.304	-0.069
t(H−L)						(-4.55)
# stocks	679	678	678	678	679	

The table reports the properties of the AI portfolios. For each formation week, firms are sorted into quintiles by their pre-formation rolling AI beta, β^{AI}. Firm betas are estimated from a 13-week rolling regression of weekly log-excess returns on the log-excess market return and the AI factor, with at least 9 non-missing weeks. The $\beta^{AI}\sigma_w(AI)$ row reports the implied portfolio return using the weekly AI-factor standard-deviation scale. Within each quintile, we take the cross-sectional average of each firm characteristic, then average across formation weeks. $\ln ME$ is log market equity. $\ln BM$ is log book-to-market. $Profit$ is gross profitability (revenues minus cost of goods scaled by assets). Inv is the year-over-year change in total assets scaled by lagged assets. Mom is the cumulative return from week $t-52$ to $t-5$. $\sigma(R)_{13w}$ is the trailing 13-week standard deviation of weekly log returns. Analyst disp. is the standard deviation of analysts' FY1 annual EPS forecasts divided by the absolute value of the mean forecast, using the latest IBES consensus available before the formation week and requiring at least two analysts. The H−L column reports the time-series average of the weekly difference between the High and Low quintile means. $t(H\text{-}L)$ is the corresponding time-series t-statistic.

momentum (Carhart, 1997) is 55.9 basis points per week; both are significant at the 5 percent level.[5]

Additional Online Appendix results further support this baseline portfolio evidence. Table IA.3 repeats the portfolio sorting exercises using each of the three input series of the AI factor ($\Delta\ln Tok$, $\Delta\ln Dol$, $\Delta\ln User$) separately and finds high-low spreads of 51.6, 36.8, and 46.3 basis points per week. Table IA.7 sorts on industry-neutral AI exposure (the difference between β^{AI} and the within-period Fama-French 30 industry median) and finds a spread of 60.3 basis points per week, close to the 64.1-basis-point baseline. Finally, to control for broad public attention and sentiment toward AI, Table IA.8 re-estimates β^{AI} controlling for the log growth of Google Trends search interest in "Artificial intelligence" and finds a spread of 69.6 basis points per week with comparable factor-model alphas.

3.2 The AI Premium

The portfolio sorts above document a return spread between high- and low-exposure firms. Table 4 further reports estimates of the market price of AI risk, the AI premium, using the Fama-MacBeth procedure

[5]Online Appendix Table IA.4 reports the factor regressions for the value-weighted, all-stock-breakpoint long-short strategy, including factor coefficient estimates and R^2 values. Online Appendix Table IA.6 shows that the result remains positive under block-bootstrap inference and when sorting on empirical-Bayes-shrunk AI betas.

Table 3: Excess Returns of AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Value-weighted, all-stock breakpoints						
Excess return	-0.002	0.280	0.293	0.510	0.639	0.641
t	(-0.01)	(1.24)	(1.34)	(2.24)	(2.18)	(2.84)
FF5 α	-0.230	0.062	0.027	0.254	0.333	0.563
$t(\alpha)$	(-1.61)	(0.65)	(0.38)	(2.84)	(2.49)	(2.43)
FF5+Mom α	-0.226	0.064	0.028	0.251	0.332	0.559
$t(\alpha)$	(-1.59)	(0.68)	(0.38)	(2.84)	(2.47)	(2.41)
Panel B. Value-weighted, NYSE breakpoints						
Excess return	0.124	0.248	0.381	0.468	0.656	0.533
t	(0.48)	(1.11)	(1.75)	(2.04)	(2.32)	(2.51)
FF5 α	-0.112	0.056	0.106	0.209	0.358	0.471
$t(\alpha)$	(-0.84)	(0.60)	(1.32)	(2.35)	(2.79)	(2.17)
FF5+Mom α	-0.108	0.058	0.106	0.206	0.358	0.466
$t(\alpha)$	(-0.82)	(0.61)	(1.32)	(2.36)	(2.77)	(2.16)
Panel C. Equal-weighted, all-stock breakpoints						
Excess return	-0.029	0.189	0.294	0.259	0.322	0.350
t	(-0.09)	(0.80)	(1.37)	(1.06)	(1.04)	(2.13)
FF5 α	-0.180	0.031	0.157	0.063	0.113	0.293
$t(\alpha)$	(-1.68)	(0.59)	(3.81)	(1.19)	(1.14)	(1.76)
FF5+Mom α	-0.176	0.034	0.157	0.062	0.111	0.287
$t(\alpha)$	(-1.68)	(0.68)	(3.79)	(1.17)	(1.12)	(1.75)
Panel D. Equal-weighted, NYSE breakpoints						
Excess return	0.013	0.221	0.294	0.260	0.322	0.309
t	(0.04)	(0.96)	(1.36)	(1.08)	(1.07)	(2.07)
FF5 α	-0.140	0.066	0.154	0.075	0.112	0.253
$t(\alpha)$	(-1.46)	(1.39)	(3.45)	(1.40)	(1.24)	(1.67)
FF5+Mom α	-0.136	0.068	0.154	0.075	0.111	0.247
$t(\alpha)$	(-1.46)	(1.49)	(3.43)	(1.38)	(1.22)	(1.66)

The table reports time-series properties of weekly returns on AI-beta quintile portfolios. Portfolios are rebalanced weekly. *Excess return* is the average post-formation-week portfolio return. *H−L* is the spread between the High and Low AI-beta quintiles. Returns and alphas are in percent per week. FF5 alphas are intercepts from regressing the portfolio's post-formation-week excess return on the Fama and French (2015) five-factor model. FF5+Mom additionally includes the momentum factor of Carhart (1997). Under *all-stock breakpoints*, the quintile cutoffs are computed from the full eligible universe. Under *NYSE breakpoints*, the quintile cutoffs are computed from NYSE-listed stocks only and applied to the full universe. Data is weekly from January 2024 through April 2026.

(Fama and MacBeth, 1973) at the individual-stock level. Each week, we regress individual stock returns on their pre-formation AI exposure, β^{AI}, and standard firm-level characteristics, and we average the weekly coefficient estimates over time. We denote the coefficient estimate on β^{AI} as the AI premium—the expected-return compensation investors require for one additional unit of market-implied AI exposure, holding other characteristics fixed. Panel A weights stocks equally, so the weekly cross-sectional regressions are estimated by OLS regressions. Panel B weights each stock by its market capitalization, so they are estimated by weighted least squares regressions.

The estimated AI premium is positive and statistically significant across specifications. When β^{AI} is the only regressor, the slope is 7.65 with a t-statistic of 2.18 in Panel A and 19.62 with a t-statistic of 3.30 in Panel B. After controlling for size, book-to-market, profitability, and investment rate, the coefficient

estimate for AI exposure remains positive and significant. The result is also robust to the most demanding specification, which adds momentum, short-term reversal, leverage, and accruals. Under this specification, the equal-weighted result is 8.82 with a t-statistic of 2.88, and the value-weighted result is 18.25 with a t-statistic of 3.58. Thus, Table 4 provides regression-based evidence that the market price of AI risk is positive.

Table 4: Fama-MacBeth Cross-Sectional Regressions

	β^{AI}	$\ln ME$	$\ln BM$	Profit	Inv	Mom	Rev	Lev	Accr	$\bar{R}^2$
Panel A. Equal-weighted										
(1)	7.65									0.006
t	(2.18)									
(2)		0.02	0.02	-0.02	-0.01					0.024
t		(0.36)	(0.25)	(-0.32)	(-0.34)					
(3)		0.02	0.04	-0.02	-0.02	0.09				0.032
t		(0.24)	(0.56)	(-0.26)	(-0.60)	(1.35)				
(4)	7.70	0.00	0.01	-0.03	-0.01					0.030
t	(2.42)	(0.07)	(0.11)	(-0.44)	(-0.29)					
(5)	7.37	-0.00	0.03	-0.03	-0.02	0.08				0.037
t	(2.38)	(-0.05)	(0.39)	(-0.39)	(-0.53)	(1.19)				
(6)		0.00	0.01	-0.01	-0.02	0.09	-0.02	-0.03	0.09	0.044
t		(0.07)	(0.23)	(-0.19)	(-0.63)	(1.30)	(-0.26)	(-0.80)	(1.67)	
(7)	8.82	-0.01	0.00	-0.02	-0.02	0.08	-0.03	-0.02	0.08	0.049
t	(2.88)	(-0.23)	(0.05)	(-0.29)	(-0.60)	(1.17)	(-0.44)	(-0.67)	(1.59)	
Panel B. Value-weighted										
(1)	19.62									0.024
t	(3.30)									
(2)		0.09	-0.04	-0.00	0.11					0.129
t		(1.12)	(-0.59)	(-0.06)	(0.87)					
(3)		0.09	-0.03	-0.01	0.09	0.05				0.166
t		(1.04)	(-0.44)	(-0.07)	(0.84)	(0.33)				
(4)	16.02	0.09	-0.06	-0.02	0.10					0.148
t	(3.13)	(1.02)	(-0.97)	(-0.23)	(0.85)					
(5)	14.30	0.08	-0.05	-0.02	0.08	0.06				0.183
t	(2.78)	(0.97)	(-0.82)	(-0.27)	(0.74)	(0.42)				
(6)		0.09	-0.04	-0.00	0.10	0.03	0.06	-0.09	0.03	0.213
t		(1.03)	(-0.69)	(-0.06)	(0.92)	(0.26)	(0.55)	(-1.48)	(0.27)	
(7)	18.25	0.09	-0.06	-0.03	0.08	0.04	0.02	-0.10	0.00	0.227
t	(3.58)	(1.02)	(-1.03)	(-0.41)	(0.78)	(0.28)	(0.21)	(-1.66)	(0.01)	

This table reports the second step of the Fama-MacBeth procedure, estimated at the individual-stock level. Each week, individual stock returns are regressed cross-sectionally on the variables shown. Panel A weights stocks equally, so the weekly regressions are estimated by OLS. Panel B weights each stock by its market capitalization, so they are estimated by weighted least squares. β^{AI} is the pre-formation rolling AI beta estimated in the first-step return regression, and its coefficient estimates the market price of AI risk, or AI premium. The remaining variables are firm-level characteristics included as controls. $\ln ME$ is log formation-week market equity, $\ln BM$ is log book-to-market, $Profit$ is gross profitability, Inv is asset growth, Mom is prior 52-to-5-week return, Rev is prior one-month return, Lev is long-term debt to assets, and $Accr$ is accruals. Characteristics are weekly cross-sectionally winsorized at the 1st/99th percentiles and standardized to unit variance. Reported coefficients are time-series averages of the weekly cross-sectional estimates. t-statistics are reported in parentheses and are time-series t-tests of those estimates (Fama and MacBeth, 1973). The $\bar{R}^2$ column is the time-series average of weekly cross-sectional R^2. Data is weekly from January 2024 through April 2026.

3.3 Additional Results

This subsection reports three sets of additional results. First, we first show that the AI premium is robust to controlling for technology-sector and AI-themed equity returns. Second, we then extend the test to international markets as a cross-market out-of-sample check. Third, we examine which subcomponents of AI consumption are most informative for asset prices.

3.3.1 Technology-Sector and AI-Themed Controls

As a robustness test, we test whether the AI premium reflects returns to technology-sector or AI-themed equity portfolios rather than information about AI consumption from OpenRouter. This is especially relevant in our short sample, a period in which technology stocks and AI-themed valuations rose sharply. We address this by re-estimating each firm's AI exposure with one of three traded benchmarks added to the market control: the high-technology industry return from Fama-French 10 industry, the value-weighted semiconductor portfolio return, and an equal-weighted AI/robotics ETF basket return. We summarize these results in Table 5. Panels A, B, and C report the results for high-low spreads for these three robustness tests. The spreads remain positive and statistically significant across the three specifications, at 49.5, 45.4, and 63.4 basis points per week, respectively.

3.3.2 International Evidence

We extend the test to international markets to ask whether the AI premium is concentrated in the markets most directly connected to leading-edge AI development, infrastructure, and commercial deployment. This builds on the "distance to frontier" framework of Acemoglu, Aghion and Zilibotti (2006), in which an economy's position relative to the global technology frontier shapes its innovation and growth dynamics. In our setting, the local equity markets whose listed firms and investors are closer to AI development and deployment are what we consider the AI frontier markets, where the AI risk is likely to be more systematic (Pastor and Veronesi, 2009). We define a local market as a country-level equity market. Using Compustat Global common stocks, each week and separately within each local market, we sort stocks into quintiles by AI exposure. We then compute the high-low return spread within each local market and average these local-market spreads across markets classified by MSCI as developed or emerging.

Panel A of Table 6 combines both local markets classified by MSCI as developed and emerging. The overall international high-low spread is 11.0 basis points per week, with a significant local-market alpha of 10.6 basis points. Panels B and C show that this average is mainly driven by developed local markets. In developed markets, the average high-low spread is 17.9 basis points per week with a 17.2-basis-point local-market alpha, both statistically significant. In emerging local markets, the average spread is 5.0 basis points per week and is statistically insignificant. The AI premium is more pronounced in developed equity markets near the global AI frontier.

China is a particularly interesting test because Chinese AI consumption has become an important share of the global total and AI is highly salient to local investors. Online Appendix Table IA.14 examines China A-shares, sorting either on the baseline AI factor or on a China-specific AI factor built from mainland-China

Table 5: High-Tech-, Semiconductor-, and AI-ETF-Controlled AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. FF10 HiTec industry as control						
Excess return	0.169	0.270	0.282	0.423	0.664	0.495
t	(0.65)	(1.21)	(1.30)	(1.79)	(2.24)	(2.49)
FF5 α	-0.058	0.030	0.041	0.142	0.343	0.401
$t(\alpha)$	(-0.42)	(0.35)	(0.59)	(1.44)	(2.71)	(2.00)
FF5+Mom α	-0.051	0.033	0.039	0.141	0.341	0.392
$t(\alpha)$	(-0.39)	(0.39)	(0.57)	(1.42)	(2.69)	(2.01)
Panel B. NAICS semiconductor industry as control						
Excess return	0.159	0.381	0.263	0.402	0.613	0.454
t	(0.57)	(1.67)	(1.15)	(1.68)	(2.25)	(2.23)
FF5 α	-0.094	0.180	-0.022	0.106	0.312	0.406
$t(\alpha)$	(-0.66)	(2.20)	(-0.27)	(1.18)	(2.66)	(1.93)
FF5+Mom α	-0.089	0.182	-0.021	0.103	0.310	0.399
$t(\alpha)$	(-0.65)	(2.22)	(-0.26)	(1.17)	(2.65)	(1.95)
Panel C. AI ETF basket as control						
Excess return	0.075	0.219	0.334	0.435	0.709	0.634
t	(0.29)	(1.01)	(1.59)	(1.81)	(2.44)	(3.11)
FF5 α	-0.150	0.003	0.072	0.167	0.394	0.544
$t(\alpha)$	(-1.18)	(0.03)	(0.88)	(1.88)	(3.00)	(2.60)
FF5+Mom α	-0.144	0.004	0.073	0.164	0.393	0.537
$t(\alpha)$	(-1.17)	(0.04)	(0.90)	(1.87)	(2.98)	(2.60)

All panels report value-weighted quintile sorts using all-stock breakpoints. Panel A expands the first-pass regression to include the Fama-French 10-industry HiTec return: $r^e_{i,t} = a + b_{\text{mkt}}\, r^e_{m,t} + b_{\text{HiTec}}\, r^e_{\text{HiTec},t} + b_{\text{AI}}\, AI_t$. Panel B instead controls for the value-weighted weekly log return on the U.S. semiconductor industry: $r^e_{i,t} = a + b_{\text{mkt}}\, r^e_{m,t} + b_{\text{semi}}\, r_{\text{semi},t} + b_{\text{AI}}\, AI_t$. Panel C controls for the equal-weighted AI ETF basket return: $r^e_{i,t} = a + b_{\text{mkt}}\, r^e_{m,t} + b_{\text{ETF}}\, r^e_{\text{ETF},t} + b_{\text{AI}}\, AI_t$. First-pass regressions require at least 9 non-missing weeks. Returns and alphas are in percent per week. Data is weekly from January 2024 through April 2026.

OpenRouter consumption. In both specifications, the high-minus-low spreads are negative and statistically insignificant.

Table 6: International AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Full international sample						
Excess return	0.207	0.271	0.257	0.260	0.317	0.110
t	(1.92)	(2.83)	(3.07)	(2.71)	(2.85)	(2.52)
Local-mkt α	0.056	0.124	0.131	0.118	0.162	0.106
$t(\alpha)$	(0.85)	(2.69)	(3.06)	(2.29)	(2.41)	(2.38)
Panel B. MSCI Developed						
Excess return	0.182	0.294	0.278	0.323	0.361	0.179
t	(1.40)	(2.38)	(2.75)	(2.87)	(2.67)	(2.77)
Local-mkt α	0.018	0.126	0.139	0.171	0.191	0.172
$t(\alpha)$	(0.23)	(2.05)	(2.85)	(2.93)	(2.35)	(2.63)
Panel C. MSCI Emerging						
Excess return	0.226	0.248	0.237	0.205	0.277	0.050
t	(2.05)	(2.75)	(2.80)	(2.06)	(2.53)	(0.94)
Local-mkt α	0.063	0.103	0.107	0.052	0.112	0.049
$t(\alpha)$	(0.79)	(1.77)	(1.83)	(0.77)	(1.45)	(0.90)

The table reports returns and local-market alphas for international pre-formation AI-beta-sorted quintile portfolios. A local market is a country-level equity market. Panel A pools local markets classified by MSCI as developed or emerging (the full international sample), and excludes frontier, standalone, and not-rated markets. Panel B restricts the sample to developed local markets and Panel C to emerging local markets. For each local market and week, stocks are sorted into quintiles using local-market breakpoints, and quintile returns are equal-weighted across firms. H−L is formed within each local market. Portfolio returns and H−L spreads are then equal-weighted across local markets. Firm betas are 13-week rolling log-excess regressions of weekly local-currency returns on the local-market log-excess return and the AI factor, with at least 9 non-missing weeks. Daily returns are JKP-style (Jensen, Kelly and Pedersen, 2023) winsorized using same-month CRSP cutoffs before weekly compounding. *Local-mkt* α is the intercept from regressing the portfolio's next-week return on the value-weighted market return of the included local markets. Data is weekly from January 2024 through April 2026.

3.3.3 Salient components of AI Consumption

The unique feature of the OpenRouter AI dataset is its granularity. For each anonymized user and model on each day, we observe realized requests, tokens, and dollar usage. This user–model–day structure lets us ask which components of AI consumption best capture priced AI risk. We decompose AI consumption along four margins: closed-source versus open-weight models, paid/core versus new accounts, seasoned versus non-seasoned users, and long versus short prompts. For each of the resulting eight specific AI factors, we re-estimate firm-level exposures and sort firms on that exposure separately. Table 7 reports the resulting high-minus-low spreads.

The four splits isolate different margins of AI consumption. First, following OpenRouter's model taxonomy, closed-source models are proprietary or restricted-API models whose weights are not publicly released, while open-weight models have publicly available weights. Second, paid/core users are mature accounts with at least 13 weeks of tenure, at least eight prior active weeks, at least $10 of prior OpenRouter usage, and activity in at least two of the previous eight weeks. New users are accounts with less than four weeks of tenure, and middle/developing accounts are excluded from this split. Third, seasoned users are a broader experience measure: before week t, they have been observed for at least 13 weeks, were active in at least four prior weeks, and made at least 20 prior requests. The remaining active users are non-seasoned. Finally, long

prompts are requests above the prior-13-week rolling 75th percentile of prompt length, and short prompts are the remaining requests. All user and prompt-length classifications are predetermined using information available before the portfolio-formation week.

The pattern is consistent across the four splits: the priced part of AI consumption is the more intensive, frontier-oriented part. The high-low spread is 53.4 basis points per week based on exposure to closed-source model consumption versus 32.3 based on open-weight consumption (Panel A), 66.8 versus 21.9 based on paid/core accounts versus new accounts (Panel B), 59.3 versus 29.3 based on seasoned versus non-seasoned users (Panel C), and 54.1 versus 33.0 based on long- versus short-prompt token growth (Panel D). The AI premium is better captured by AI consumption on frontier models, among experienced users, and for complex tasks, rather than on the casual or extensive margin.

Table 7: AI-Beta Quintile Portfolios by Salient Consumption Components

	Low	2	3	4	High	H−L
Panel A. AI model class						
$\beta^{AI,Closed}$						
Excess return	0.144	0.156	0.431	0.381	0.678	0.534
t	(0.53)	(0.76)	(1.94)	(1.61)	(2.29)	(2.47)
FF5 α	-0.095	-0.053	0.187	0.090	0.338	0.433
$t(\alpha)$	(-0.64)	(-0.66)	(2.36)	(1.07)	(2.87)	(1.96)
FF5+Mom α	-0.091	-0.053	0.188	0.088	0.337	0.428
$t(\alpha)$	(-0.62)	(-0.65)	(2.38)	(1.05)	(2.84)	(1.94)
$\beta^{AI,Open}$						
Excess return	0.225	0.316	0.442	0.285	0.547	0.323
t	(0.81)	(1.43)	(2.26)	(1.26)	(1.85)	(1.76)
FF5 α	-0.060	0.043	0.201	0.044	0.224	0.284
$t(\alpha)$	(-0.49)	(0.62)	(2.61)	(0.54)	(1.92)	(1.49)
FF5+Mom α	-0.057	0.043	0.200	0.044	0.224	0.281
$t(\alpha)$	(-0.46)	(0.61)	(2.59)	(0.54)	(1.91)	(1.47)
Panel B. Paid/core versus new users						
$\beta^{AI,Core}$						
Excess return	0.068	0.266	0.348	0.473	0.736	0.668
t	(0.25)	(1.16)	(1.50)	(1.67)	(1.97)	(2.37)
FF5 α	-0.180	0.079	0.105	0.151	0.342	0.522
$t(\alpha)$	(-1.13)	(0.82)	(1.13)	(1.60)	(2.10)	(1.95)
FF5+Mom α	-0.173	0.077	0.104	0.150	0.341	0.514
$t(\alpha)$	(-1.14)	(0.81)	(1.12)	(1.58)	(2.08)	(1.95)
$\beta^{AI,New}$						
Excess return	0.299	0.420	0.362	0.302	0.518	0.219
t	(0.89)	(1.71)	(1.53)	(1.24)	(1.61)	(0.86)
FF5 α	-0.078	0.207	0.118	0.036	0.242	0.320
$t(\alpha)$	(-0.53)	(1.89)	(1.35)	(0.38)	(1.66)	(1.35)
FF5+Mom α	-0.076	0.208	0.118	0.037	0.241	0.318
$t(\alpha)$	(-0.51)	(1.91)	(1.33)	(0.39)	(1.65)	(1.34)
Panel C. Seasoned versus non-seasoned users						
$\beta^{AI,Seasoned}$						
Excess return	0.122	0.183	0.429	0.421	0.715	0.593
t	(0.43)	(0.83)	(1.79)	(1.59)	(1.88)	(2.13)
FF5 α	-0.159	-0.001	0.199	0.099	0.307	0.466

	Low	2	3	4	High	H–L
$t(\alpha)$	(-0.97)	(-0.02)	(2.04)	(1.19)	(1.81)	(1.73)
FF5+Mom α	-0.155	-0.001	0.197	0.097	0.308	0.462
$t(\alpha)$	(-0.95)	(-0.01)	(2.03)	(1.17)	(1.80)	(1.71)
$\beta^{AI,Nonseasoned}$						
Excess return	0.207	0.519	0.214	0.326	0.500	0.293
t	(0.54)	(2.09)	(0.88)	(1.38)	(1.62)	(1.21)
FF5 α	-0.123	0.280	-0.020	0.075	0.155	0.278
$t(\alpha)$	(-0.72)	(2.63)	(-0.23)	(0.88)	(1.32)	(1.22)
FF5+Mom α	-0.123	0.286	-0.020	0.074	0.153	0.276
$t(\alpha)$	(-0.71)	(2.85)	(-0.23)	(0.87)	(1.30)	(1.21)
Panel D. Prompt length						
$\beta^{Tok,Long}$						
Excess return	0.085	0.206	0.454	0.327	0.626	0.541
t	(0.33)	(0.95)	(2.01)	(1.33)	(2.07)	(2.40)
FF5 α	-0.150	-0.008	0.188	0.050	0.278	0.428
$t(\alpha)$	(-1.05)	(-0.09)	(1.91)	(0.56)	(2.14)	(1.91)
FF5+Mom α	-0.146	-0.005	0.189	0.048	0.274	0.420
$t(\alpha)$	(-1.05)	(-0.06)	(1.93)	(0.55)	(2.15)	(1.94)
$\beta^{Tok,Short}$						
Excess return	0.325	0.321	0.341	0.274	0.655	0.330
t	(1.12)	(1.41)	(1.59)	(1.25)	(2.19)	(1.79)
FF5 α	0.048	0.022	0.042	0.024	0.365	0.318
$t(\alpha)$	(0.40)	(0.28)	(0.56)	(0.29)	(2.61)	(1.68)
FF5+Mom α	0.049	0.023	0.043	0.023	0.366	0.317
$t(\alpha)$	(0.41)	(0.29)	(0.56)	(0.27)	(2.60)	(1.67)

This table reports value-weighted, all-stock-breakpoint quintile sorts on split-specific AI exposures. Panels A–D compare closed- and open-weight model consumption, paid/core and new accounts, seasoned and non-seasoned users, and long- and short-prompt token growth, respectively (see Section 3.3.3 for the definitions of the groups). Rolling beta regressions require at least 9 non-missing weeks. Returns and alphas are in percent per week. Data is weekly from January 2024 through April 2026.

3.3.4 Event-Study Evidence

AI is a recent phenomenon, and a series of unanticipated positive surprises could, in principle, explain part of the observed average excess return. If high-AI-beta firms are those who benefit most as AI advances, then AI advancing faster than the market expected would also be consistent with high realized returns over our sample. We provide two complementary robustness results to investigate this possibility: an event study around AI model releases, and an analysis that removes those release dates from the sample.

First, we use major AI model releases as the news events, separating releases by the five frontier providers—Anthropic, DeepSeek, Google, Meta, and OpenAI—from those of other providers, and compare the cumulative abnormal returns (CARs) of high- and low-AI-beta firms in narrow windows around release dates. Our constructed release calendar, containing 19 frontier and 28 non-frontier events, is reported in the Online Appendix (Table IA.16). Figure 2 plots the high-minus-low CAR difference from five days before to ten days after each release, with 95 percent confidence intervals. The effect is concentrated in frontier-provider releases (Panel A): the CAR difference reaches about 3 percent, with evidence of pre-drift before the release and a plateau about five days after, and is statistically significant within the window. For

non-frontier releases (Panel B) it reaches only about 1.5 percent, with little pre-drift. These results are consistent with high-AI-beta firms loading on a common, systematic AI factor and frontier releases—the most informative events—moving them the most. We report the full set of risk-adjusted CAR differences across event windows in the Online Appendix (Table IA.15).

Second, we remove the frontier model-release weeks from the sample. If the premium were only the ex-post realization of AI news, it should be concentrated in those weeks, and excluding them should eliminate it. It does not. Table 8 repeats the value-weighted quintile sort with these weeks excluded, identified from the model-release calendar in Table IA.16. Dropping weeks with a frontier-provider release, the high-minus-low spread remains positive at 0.395 percent per week ($t = 1.96$), and additionally dropping non-frontier-release weeks leaves 0.456 percent per week ($t = 1.77$). In both cases, the high AI-beta quintile has an higher average return than the low quintile. The premium is stronger around releases, as the event study shows, but it accrues gradually throughout the sample rather than only on news dates.

Table 8: AI-Beta Quintile Portfolios Excluding Model-Release Weeks

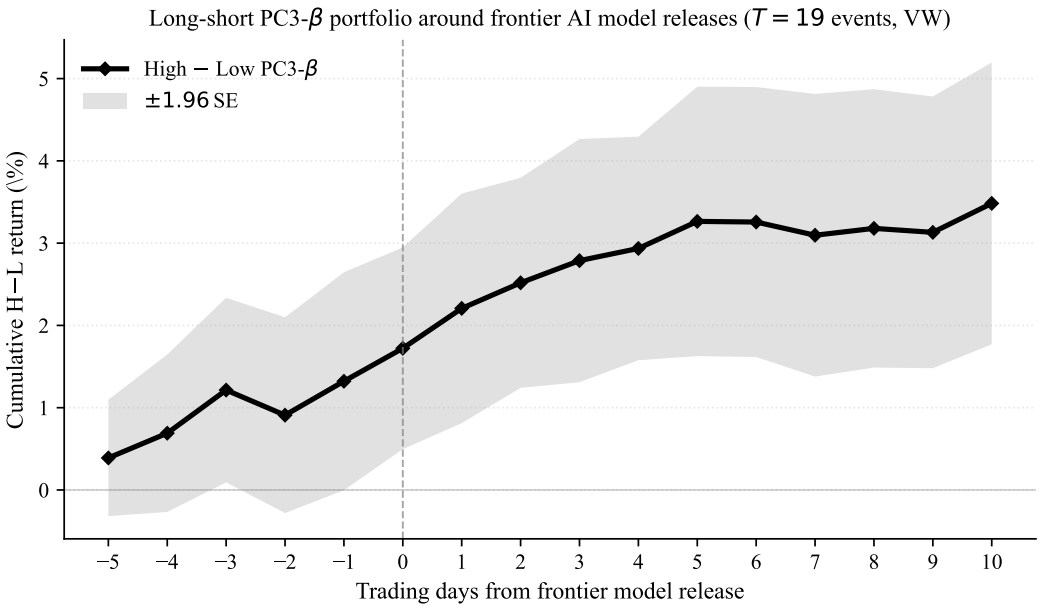
	Low	2	3	4	High	H−L
Panel A. Excluding frontier-release weeks (value-weighted, all-stock breakpoints)						
Excess return	0.160	0.306	0.282	0.423**	0.555**	0.395**
t	(0.66)	(1.45)	(1.31)	(2.02)	(2.11)	(1.96)
Panel B. Excluding frontier and non-frontier release weeks (value-weighted, all-stock breakpoints)						
Excess return	-0.072	0.158	0.146	0.306	0.384	0.456*
t	(-0.25)	(0.62)	(0.58)	(1.26)	(1.11)	(1.77)

The table reports weekly value-weighted excess returns on AI-beta quintile portfolios, all-stock breakpoints, after excluding weeks that contain major AI model releases. Panel A excludes holding weeks with a frontier-provider release (Anthropic, DeepSeek, Google, Meta, OpenAI); Panel B additionally excludes weeks with releases from other providers. The sorting signal, universe, and portfolio construction match Table 3. *H−L* buys the High and shorts the Low AI-beta quintile. Returns are in percent per week; Newey–West *t*-statistics with three lags are in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent levels.

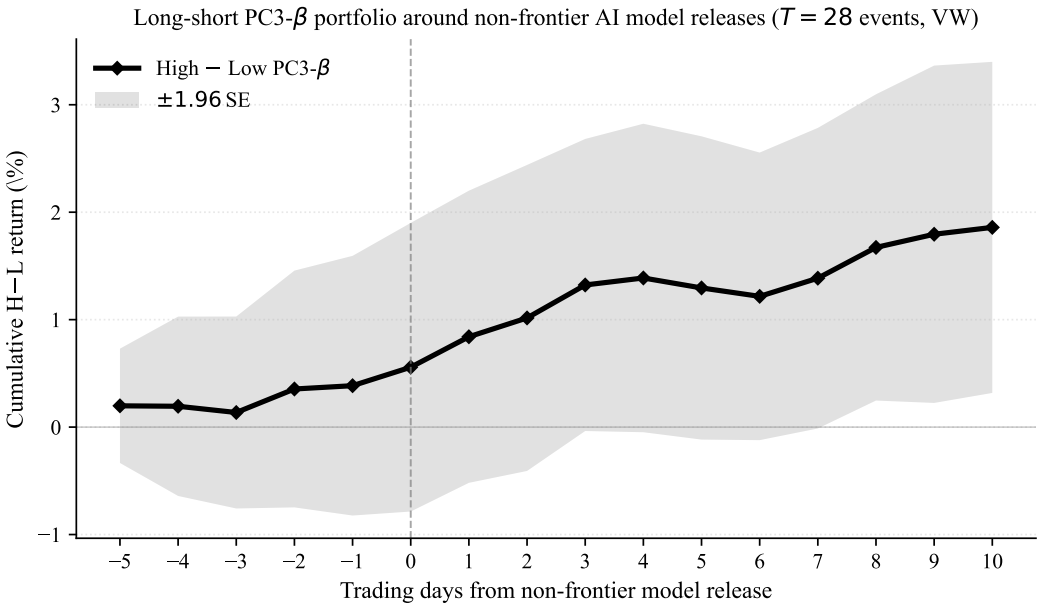
3.4 Firm-Level AI Exposure

The rolling regressions used in the portfolio tests also give us a firm-level measure of AI exposure. For each firm, we estimate β^{AI} from weekly stock returns and then average the estimated exposure across valid formation weeks. This allows us to move beyond portfolios and ask which individual firms' returns comove most strongly with innovations to AI consumption.

Figure 3 illustrates this firm-level exposure for the S&P 500 firms. Panel A reports the 50 firms with the largest positive β^{AI}, while Panel B reports the 50 firms with the largest negative β^{AI}. The figure should be read as follows. Each company is placed on a radial scale centered at zero. Distance from the center is the stock return, in percentage points per week, associated with a one-standard-deviation increase in the AI factor. In Panel A, firms farther from the center have a larger positive response to an AI-factor shock. In Panel B, firms farther from the center have a larger negative response. The two panels use the same radial scale, so distances from the center are comparable across positive and negative exposures. Label colors denote the top 3 common industries in the graph from the Fama-French 10 industries, as shown in the legend.



(A) Frontier Model Releases (19 Events)



(B) Non-Frontier Model Releases (28 Events)

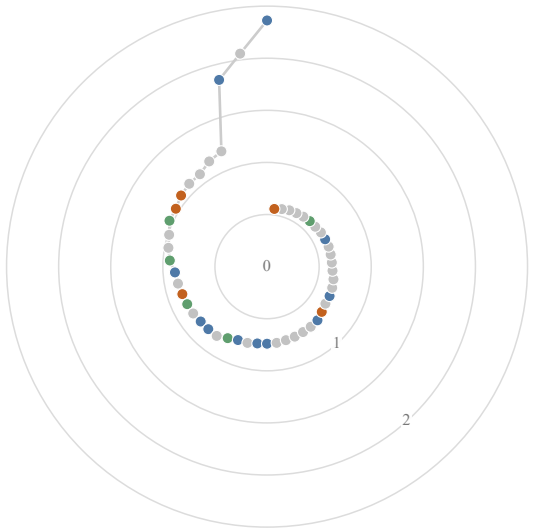
Figure 2: Cumulative Average Returns Around AI Model Releases

Daily event-time CAARs to the value-weighted, all-stock-breakpoint long-short AI-β quintile portfolio. In each panel, firms are sorted on their pre-event 13-week rolling log-excess AI beta and the H−L spread is the High-quintile minus Low-quintile cumulative return. Time is measured in trading days from the event anchor (same-or-next trading day on or after the announced release date). The shaded band is the ±1.96 cross-event standard error of the cumulated H−L return at each event-time day. Panel A uses the 19 frontier-model events of Table IA.15 Panel A; Panel B uses the 28 non-frontier-model events of Table IA.15 Panel B. Both panels are value-weighted. Data is weekly from January 2024 through April 2026.

The highest-ranked firm in Panel A, AppLovin, a software company whose platform helps mobile-app developers advertise and monetize apps, has the highest estimated positive AI exposure among current S&P 500 firms. The highest-ranked firm in Panel B, Moderna, a biotechnology firm focused on mRNA medicines and vaccines, has the most negative estimated AI exposure. The positive tail also includes natural AI-infrastructure names such as NVIDIA and Lumentum, digital-platform and travel firms such as Expedia, and power firms such as NRG Energy. The negative tail includes Estee Lauder, as well as technology and semiconductor firms such as AMD and ON Semiconductor. The main point of this result is that firm-level AI exposure is not the same as a simple technology-sector classification, our AI exposure measure identifies firms whose returns covary with AI consumption across many parts of the economy.

Panel A. Top 50 S&P 500 Firms

1. APP AppLovin
2. CVNA Carvana
3. LITE Lumentum
4. EXE Expand Energy
5. BKR Baker Hughes
6. EXPE Expedia Group
7. TPR Tapestry, Inc.
8. NRG NRG Energy
9. GEV GE Vernova
10. CMI Cummins
11. NCLH Norwegian Cruise Line Holdings
12. PODD Insulet Corporation
13. JCI Johnson Controls
14. FFIV F5, Inc.
15. PGR Progressive Corporation
16. TRGP Targa Resources
17. PH Parker Hannifin
18. EQT EQT Corporation
19. WDC Western Digital
20. COHR Coherent Corp.
21. C Citigroup
22. KVUE Kenvue
23. CIEN Ciena
24. RL Ralph Lauren Corporation
25. TRMB Trimble Inc.
26. GRMN Garmin
27. RF Regions Financial Corporation
28. SATS EchoStar
29. CFG Citizens Financial Group
30. UAL United Airlines Holdings
31. KEY KeyCorp
32. STX Seagate Technology
33. OKE Oneok
34. WM Waste Management
35. ADSK Autodesk
36. IBKR Interactive Brokers
37. ORLY O'Reilly Automotive
38. LYV Live Nation Entertainment
39. AON Aon plc
40. CTAS Cintas
41. SCHW Charles Schwab Corporation
42. INTU Intuit
43. FIX Comfort Systems USA
44. FCX Freeport-McMoRan
45. GLW Corning Inc.
46. TAP Molson Coors Beverage Company
47. VMC Vulcan Materials Company
48. ROL Rollins, Inc.
49. AJG Arthur J. Gallagher & Co.
50. ETR Entergy



High-tech Manufacturing Utilities All other industries

Figure 3: S&P 500 Firm-Level AI Exposure

Panel B. Bottom 50 S&P 500 Firms

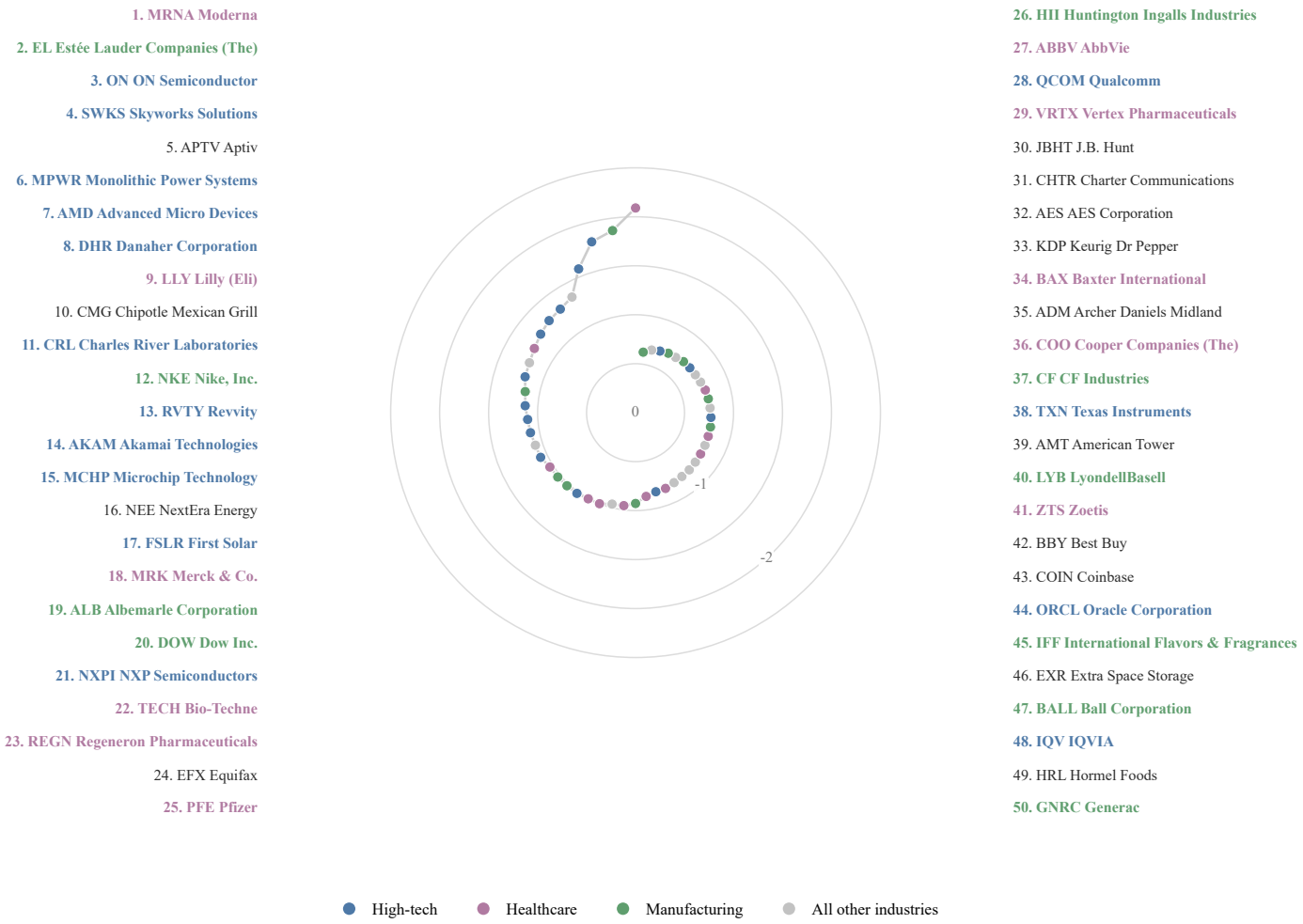


Figure 3: S&P 500 Firm-Level AI Exposure (continued)

The figure plots the 50 current S&P 500 firms with the largest positive (Panel A) and most negative (Panel B) average rolling exposures to the baseline AI factor. Firm-level β^{AI} is estimated from 13-week rolling regressions of weekly log-excess returns on the market log-excess return and the AI factor, then averaged across valid weeks. Radial distance reports $100\times\overline{\beta}_i^{AI}\times\sigma_w(AI)$, the weekly return in percentage points associated with a one-standard-deviation AI-factor shock. Label colors denote Fama-French 10 industries, as shown in the legend. Both panels use the same radial scale.

4 AI Exposure across Occupations, Tasks and Skills

This section provides the occupation-level, task-level, and skill-level counterparts of the firm-level results. We use the firm-level AI betas to map market-priced AI exposure to occupations, tasks and skills in the current organization of work. We proceed in two steps. First, we map the firm-level AI betas into occupations using BLS occupation-by-industry employment weights (the occupation ranking is in Online Appendix Figure IA.6). Second, we map these occupation exposures into skills using O*NET, the U.S. Department of Labor database that describes the tasks, abilities, and skills used in detailed occupations, or the skill-importance ratings for each occupation.

The firm-level AI exposure comes directly from stock returns and AI-consumption innovations as described above, the occupation mapping adds employment weights, and the skill mapping adds O*NET skill-importance ratings. In the end, this procedure produces a market-based measure of AI exposure for occupations and skills. This measure is thus different from worker surveys, expert labels, or AI labels which ask AI models to judge which skills are exposed. This is the first exercise that uses realized AI consumption and equity market valuations to build a market-implied occupation and skill-level measure of AI exposure.

Figure 4 reports the skill-level AI exposure ranking. Panel A uses the current occupation-by-industry employment structure, and Panel B repeats the exercise with BLS projected 2034 occupation weights. The ranking is similar across the two panels. Positive AI exposure is concentrated in installation, repair, programming, persuasion, instructing, and systems-oriented skills, and negative AI exposure in science and operations-control skills. The figure shows which skills are more prominent in the firms and industries whose market values are most sensitive to AI. In other words, the figure tells us, conditional on AI's future success, the skills the market expects to gain the most and the skills to lose the most.

Table 9 compares the same market-implied occupation exposures with task and skill measures from the literature on technological change and the demand for labor. Tasks are the activities a job involves, and skills are the worker capabilities those activities require. Columns (1) and (2) of the table use task-content measures. Column (1) follows the work of Autor, Levy and Murnane (2003) and assigns each occupation four scores: routine, nonroutine analytic, nonroutine interactive, and manual/physical. Column (2) keeps the two nonroutine cognitive measures and splits the rest into routine cognitive, routine manual, and nonroutine manual, following the five-group taxonomy of Acemoglu and Autor (2011). Columns (3) through (5) assign the skill-based measures of Deming (2017) to each occupation, including the three headline measures in column (3) (social skills, nonroutine analytical and mathematical content, and routine tasks), the interaction between social skills and analytical content that captures his social-cognitive complementarity in column (4), and the full set of ten O*NET task and skill measures in column (5). All variables are standardized, so coefficients are in standard-deviation units.

We find a similar pattern across these different task and skill taxonomies, from the routine versus nonroutine split of Autor, Levy and Murnane (2003) and Acemoglu and Autor (2011) to the skill-based measures of Deming (2017). Market-implied AI exposure is positive on nonroutine interactive tasks and social skills, and negative on nonroutine analytical and math tasks. In column (1), nonroutine interactive content enters at 0.15 ($t = 2.87$) and nonroutine analytic content at -0.15 ($t = -3.41$). The Deming measures in column (3)

show the same split, 0.16 for social skills and −0.13 for nonroutine analytical and math, both significant at the 1 percent level. The composite routine and manual measures in column (1) are small and insignificant, but splitting them in column (2) shows two offsetting margins: routine manual content has a strongly negative coefficient (−0.41) and nonroutine manual content a positive one (0.25), in line with the installation and repair skills near the top of Figure 4, and the two nonroutine cognitive coefficients keep their signs and significance. Column (4) shows that the interaction between social skills and analytical and math content is small and insignificant: AI exposure is associated with social skills directly, not with the social-cognitive combination that Deming (2017) emphasizes. Column (5) splits the composite measures into ten task and skill measures and shows where the pattern comes from in detail. The positive coefficient comes from interaction/communication tasks (0.36, $t = 4.21$), and the negative coefficients from deductive and inductive reasoning and from information use. That is, with this finer measure, the composite analytical and math measure of an occupation is no longer significantly associated with the AI exposures of the occupation. Service tasks also have a negative coefficient, in line with the negative AI exposure of healthcare occupations in Online Appendix Figure IA.6.

The positive coefficients on nonroutine interactive tasks and social skills indicate that market-implied AI exposure is concentrated in occupations built around communication, persuasion, instruction, and coordination. The negative coefficients on nonroutine analytical and math content, deductive and inductive reasoning, and information use indicate that this AI exposure is not only a measure of abstract analytical or scientific work. If anything, the occupations most intensive in those tasks have the most negative AI exposure. The explanatory power is modest, where task and skill contents explain between 2 and 11 percent of the cross-sectional variation in market-implied AI exposure. Jones and Tonetti (2026) build on the task framework and emphasize that output is constrained by tasks where labor improves only slowly relative to capital—the "weak links" that bind even when most other tasks are automated. The positively-exposed skills in our map—interaction, persuasion, instruction, installation and repair—maybe the weak links the market expects to bind production as AI automates other tasks.

Table 10 reports regressions of the market-implied ranking on existing occupation-level measures of AI exposure. Column (1) uses three technical measures: the ability-based AI occupational exposure of Felten, Raj and Seamans (2021), the GPT-4 task-exposure measure of Eloundou, Manning, Mishkin and Rock (2023), and the patent-based measure of Webb (2020). Column (2) uses the generative-AI task exposure of Eisfeldt, Schubert and Zhang (2023), split into exposure through core tasks, which they tie to labor substitution, and exposure through supplemental tasks, which they tie to complementarity. Column (3) combines the five measures. Two results stand out. First, market-implied AI exposure is largely unrelated to all of these measures: its correlation with each is below 0.08 in absolute value, and no specification explains more than 2 percent of the cross-sectional variation. Second, our measure shows some correlation with the recent measures built using current generative AI models, but not with the measures that predate them. The coefficient estimate to the GPT-4 exposure measure is positive at 0.16 ($t = 2.72$), and that to the Eisfeldt, Schubert and Zhang (2023) measures is also positive, with the supplemental-task margin marginally significant. The two pre-generative-AI measures do not line up: the coefficient estimate to the Felten, Raj and Seamans (2021) measure is insignificant, and the coefficient estimate to the Webb measure is significantly negative. In the

combined column, the Felten, Raj and Seamans (2021), GPT-4, and Eisfeldt, Schubert and Zhang (2023) measures draw on overlapping assessments of cognitive task content and are highly correlated (the GPT-4 measure correlates about 0.85 with both the Felten, Raj and Seamans (2021) and core-task measures), so the coefficient esitmates estimates become individually insignificant, while the coefficient estimate to the Webb (2020) measure stays significantly negative.

5 The Rise of Agentic AI Consumption

This section presents early evidence on the rise of agentic AI consumption—models executing multi-step workflows by planning, calling external tools, taking actions, and coordinating with other agents (see, e.g., Ng, 2024a,b). Tool use is especially central for this measurement—a model becomes agentic when it is given functions it can request to call in order to gather information, manipulate data, or take action. We measure the agentic component of AI consumption in the OpenRouter data as requests whose normalized finish reason is `tool_calls`, in which the model invoked an external tool.

These agentic requests have grown rapidly. Figure 5 plots weekly agentic and non-agentic token volumes from January 2024 through April 2026. Both series scale up by several orders of magnitude over the sample, but agentic tokens grow faster. The agentic share is small in 2024, rises sharply through 2025 and into 2026, and reaches 52.2 percent of total tokens by the latest full week (see also Yang, Yonack, Zyskowski, Yarats, Ho and Ma (2025)).

Figure 6 plots agentic token volume against the realized dollars per agentic token over the same period. Volume rises by orders of magnitude, while the realized dollars per token decline at the same time. Two forces compress the per-token price: prompt caching serves prefix tokens without recomputing them through the model, and providers route agentic requests toward cheaper backing models. Both forces reduce the revenue earned on each agentic token even as agentic usage expands.

Figure 7 brings these two patterns together. Panel A plots the agentic share of total tokens against its share of total usage dollars. The token share rises much faster than the dollar share, because volume growth (Figure 5) outpaces the per-token price decline (Figure 6). Panel B plots three token-level salient components of agentic activity, each as a share of total tokens. The tool-call share counts tokens on requests that call an external tool. The cache-read share counts prompt prefixes the platform serves from cache rather than recomputing through the model. The reasoning share counts the model's internal work plan tokens. A single agentic exchange can require all three: asked "What is the temperature in Rome?", the model produces reasoning tokens to lay out the steps, issues a tool call to an external weather API, and reads the recurring system prompt from cache rather than re-processing it. The tool-call and cache-read shares are near zero in 2024 and rise to roughly two-fifths to one-half of all tokens by the end of the sample, while the reasoning share rises far more slowly.

We then test whether exposure to these agentic-consumption components is priced in the cross-section of equity returns—whether firms whose returns covary more with these components earn higher average returns. For each factor we constructed above—agentic, cache-read, and reasoning tokens, agentic dollars, and their first principal component—we estimate firm-level betas from rolling weekly return regressions,

Table 9: Market-Implied Occupation Exposure and Task and Skill Measures

	Task content		Skill content		
	(1) ALM	(2) Acemoglu-Autor	(3) Deming	(4) + interaction	(5) detailed
Routine	-0.019				
	(-0.44)				
Nonroutine Analytic	-0.152***	-0.203***			
	(-3.41)	(-4.07)			
Nonroutine Interactive	0.149***	0.095**			
	(2.87)	(2.07)			
Manual/Physical	-0.064				
	(-1.55)				
Routine Cognitive		0.027			
		(0.56)			
Routine Manual		-0.413***			
		(-4.48)			
Nonroutine Manual		0.252***			
		(3.23)			
Social Skills			0.160***	0.155***	0.260***
			(3.44)	(3.31)	(3.74)
Nonroutine Analytical / Math			-0.125***	-0.124***	0.022
			(-3.46)	(-3.43)	(0.15)
Routine			0.065	0.062	0.116*
			(1.21)	(1.13)	(1.76)
Social Skills × Analytical / Math				-0.026	
				(-0.65)	
Service					-0.186***
					(-3.95)
Deductive / Inductive Reasoning					-0.267***
					(-3.00)
Number Facility					-0.019
					(-0.14)
Information Use					-0.288**
					(-2.43)
Required Social Interaction					0.035
					(0.63)
Coordination / Team Building					0.048
					(0.63)
Interaction / Communication					0.364***
					(4.21)
Observations	649	649	649	649	649
R^2	0.024	0.070	0.023	0.023	0.106

This table reports regressions of standardized market-implied occupation AI exposure on task and skill measures from the labor literature. Column (1) uses four Autor-Levy-Murnane (ALM) style task groups constructed from O*NET descriptors, following Autor, Levy and Murnane (2003): routine, nonroutine analytic, nonroutine interactive, and manual/physical. Column (2) uses an expanded Acemoglu-Autor style taxonomy that keeps the same nonroutine analytic and nonroutine interactive measures and splits the routine and manual margins into routine cognitive, routine manual, and nonroutine manual. Columns (1) and (2) use task-content measures. Columns (3)–(5) use the skill-based measures emphasized in Deming (2017). Column (3) uses his three headline measures: social skills, nonroutine analytical and mathematical content, and routine tasks. Column (4) adds the interaction between standardized social skills and standardized nonroutine analytical/math, following Deming's social-cognitive complementarity. Column (5) uses the ten Deming-style O*NET task and skill measures. The first three have the same definitions as in columns (3) and (4). All right-hand-side variables are standardized. The interaction is the product of standardized components. t-statistics based on HC1 robust standard errors are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Underlying firm-level AI betas are weekly from January 2024 through April 2026.

Table 10: Market-Implied Occupation Exposure and AI-Exposure Measures

	(1) Technical AI	(2) ESZ (core & suppl.)	(3) Combined
Felten AIOE	-0.079		-0.056
	(-1.26)		(-0.75)
Eloundou GPT-4	0.158***		0.125
	(2.72)		(1.46)
Webb AI	-0.078**		-0.103***
	(-2.11)		(-2.58)
ESZ GenAI Exposure (Core Tasks)		0.050	0.023
		(1.41)	(0.31)
ESZ GenAI Exposure (Supplemental Tasks)		0.045*	0.012
		(1.69)	(0.34)
Observations	649	602	602
R^2	0.015	0.006	0.018

This table reports regressions of standardized market-implied occupation AI exposure on occupation-level AI-exposure measures. Column (1) uses technical AI exposure measures from Felten, Raj and Seamans (2021), Eloundou, Manning, Mishkin and Rock (2023), and Webb (2020). Column (2) uses the occupation-level generative-AI task exposure of Eisfeldt, Schubert and Zhang (2023) (ESZ), from their public data repository: exposure through core tasks and exposure through supplemental tasks, whose sum is the occupation's total exposed task share. In ESZ, exposure through core tasks proxies for labor substitution and exposure through supplemental tasks for complementarity. Column (3) combines the technical AI and ESZ measures. ESZ scores are mapped from 2010 SOC to 2018 SOC occupations using the Census occupation-code crosswalk, which reduces the sample in columns (2) and (3). All right-hand-side variables are standardized. t-statistics based on HC1 robust standard errors are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Underlying firm-level AI betas are weekly from January 2024 through April 2026.

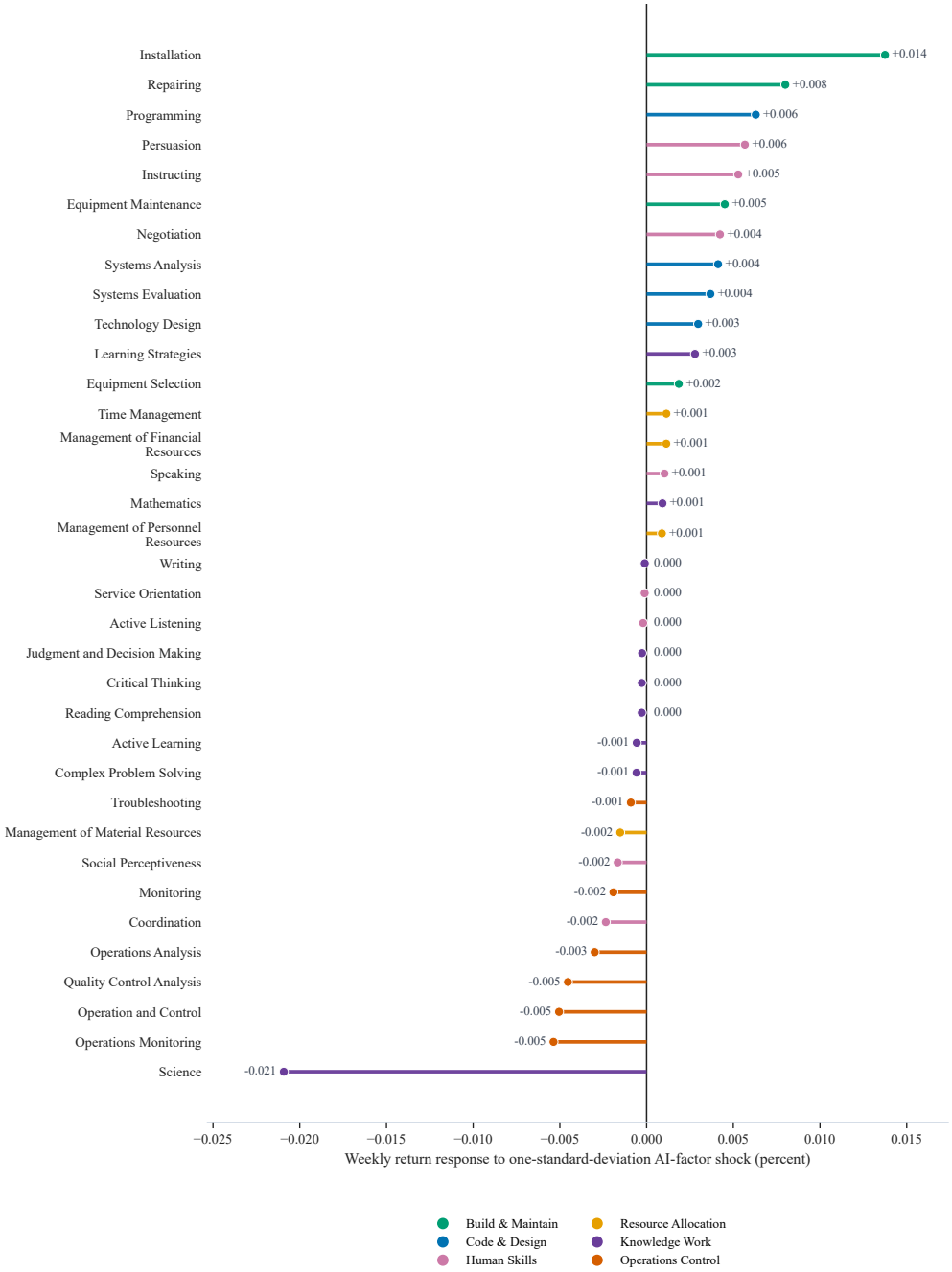
sort firms into quintiles, and compute the value-weighted high-minus-low spread. Table 11 reports these spreads, raw and as alphas from the Fama–French three-factor model (Fama and French, 1993), the Fama–French five-factor model (Fama and French, 2015), and the five-factor model augmented with momentum (Carhart, 1997). The point estimates are positive and sizable for the individual agentic factors, on the order of 0.3 to 0.5 percent per week and of the same sign as the overall AI premium. They are, however, estimated imprecisely because of the short sample. We thus take this as early evidence of a positive agentic premium.

Table 11: Cross-Sectional Pricing of Agentic Factors with Fama–French (Fama and French, 1993, 2015) and Momentum (Carhart, 1997) Adjustments

	Raw		FF3		FF5		FF5+Mom	
Factor	α	t	α	t	α	t	α	t
Agentic tokens	0.310	(0.76)	0.163	(0.51)	0.295	(0.99)	0.266	(0.90)
Cache-read tokens	0.361	(0.85)	0.337	(0.72)	0.406	(0.73)	0.520	(1.21)
Reasoning tokens	0.366	(1.58)	0.348	(1.50)	0.271	(1.15)	0.310	(1.34)
Agentic dollars	0.478	(1.18)	0.265	(0.91)	0.326	(1.07)	0.323	(1.05)
Composite (PC1)	0.161	(0.30)	0.247	(0.45)	-0.035	(-0.06)	0.058	(0.10)

The table reports value-weighted high-minus-low quintile spreads, in percent per week, for salient components of agentic AI token consumption factors; for details see Table IA.17 in the Online Appendix. Firm betas are estimated from a 13-week rolling regression of weekly log-excess returns on the log-excess market return and the factor, with a minimum of nine non-missing weeks; stocks are sorted into quintiles by this beta. Columns report the raw spread and alphas from the Fama–French three-factor model (FF3), the Fama–French five-factor model (FF5), and the five-factor model augmented with momentum (FF5+Mom), all value-weighted under all-stock breakpoints. t-statistics in parentheses.

Panel A. Current Occupation Weights



Panel B. BLS Projected 2034 Occupation Weights

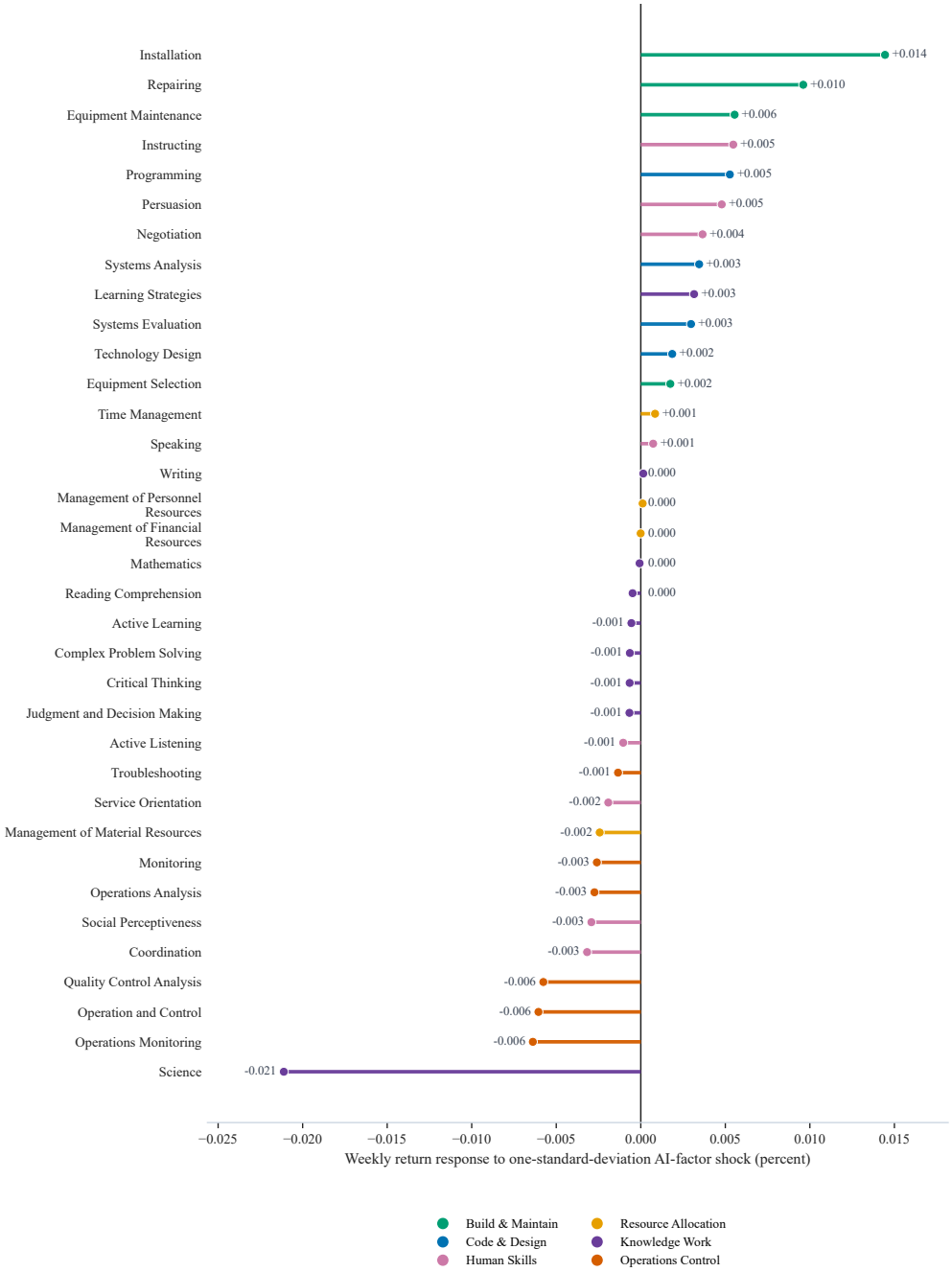


Figure 4: Skill Exposure to the AI Factor

The figure plots O*NET skill exposure to the baseline AI factor. Skill exposure is constructed by mapping detailed SOC occupation exposures to O*NET skill importance ratings and taking matched-employment-weighted averages across occupations. Panel A uses the current occupation-by-industry employment structure. Panel B uses BLS projected 2034 occupation-by-industry employment weights while holding industry AI exposures fixed. This projection is not an AI-specific displacement counterfactual. It only tests whether the static ranking is sensitive to replacing current staffing weights with BLS projected staffing weights. The plotted value is the weekly return response to a one-standard-deviation AI-factor shock, in percent. Colors group skills into six economic themes: Build & Maintain, Code & Design, Human Skills, Resource Allocation, Knowledge Work, and Operations Control. Underlying firm-level AI betas are weekly from January 2024 through April 2026.

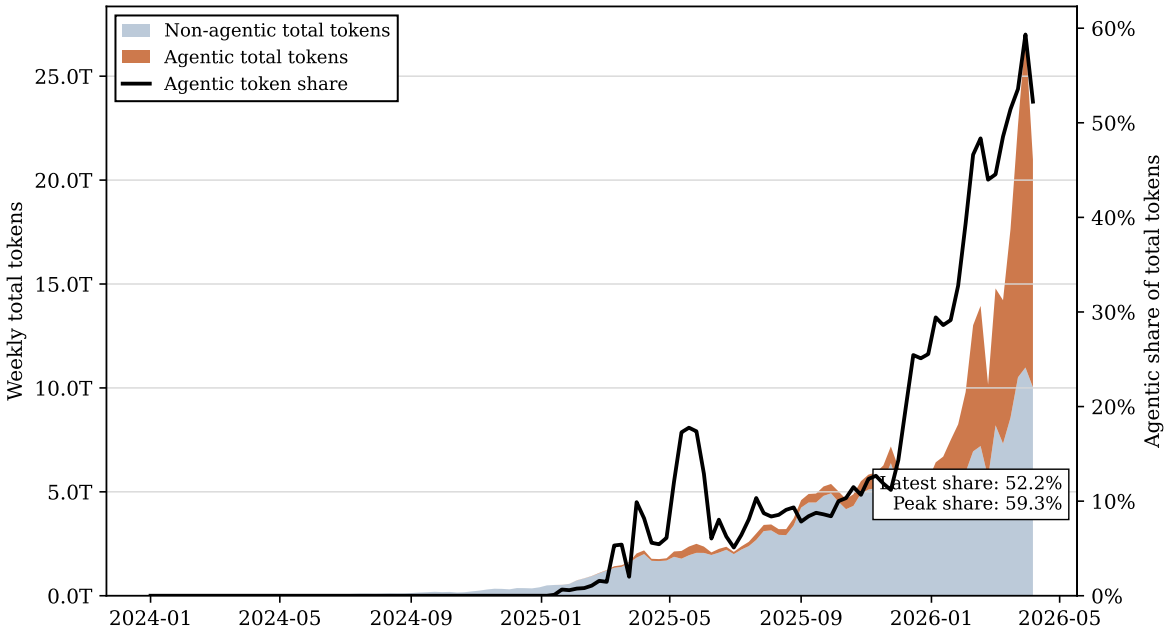


Figure 5: Agentic and Non-Agentic Token Consumption

The figure plots weekly total token consumption in the OpenRouter dataset, decomposed into agentic and non-agentic total tokens. Total tokens are defined as prompt tokens plus completion tokens. Agentic total tokens are total tokens from requests whose normalized finish reason is `tool_calls`; non-agentic total tokens are the remaining total tokens. The black line plots the agentic share of total tokens. Data is weekly from January 2024 through April 2026.

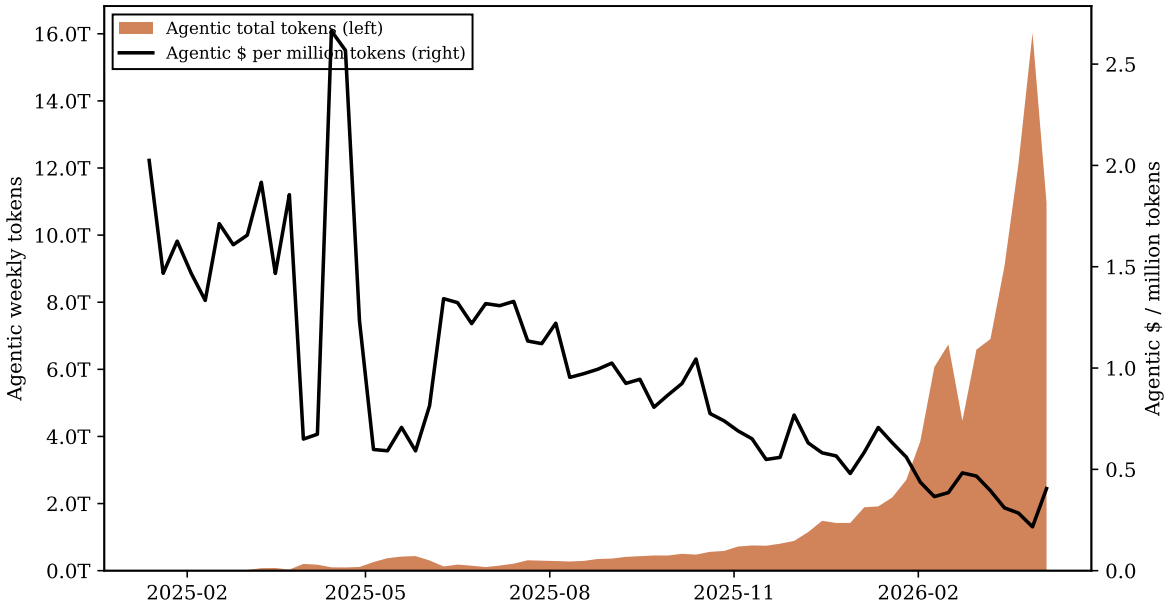


Figure 6: Agentic Token Volume and Realized Price per Token

The figure plots weekly agentic total tokens (left axis; prompt plus completion tokens from requests with normalized finish reason `tool_calls`) against the realized dollars per million tokens on the same agentic requests (right axis; OpenRouter usage dollars divided by agentic total tokens, times one million). Agentic token volume rises by orders of magnitude while the realized price per token declines, consistent with prompt caching and routing toward cheaper models compressing revenue per token even as agentic usage expands.

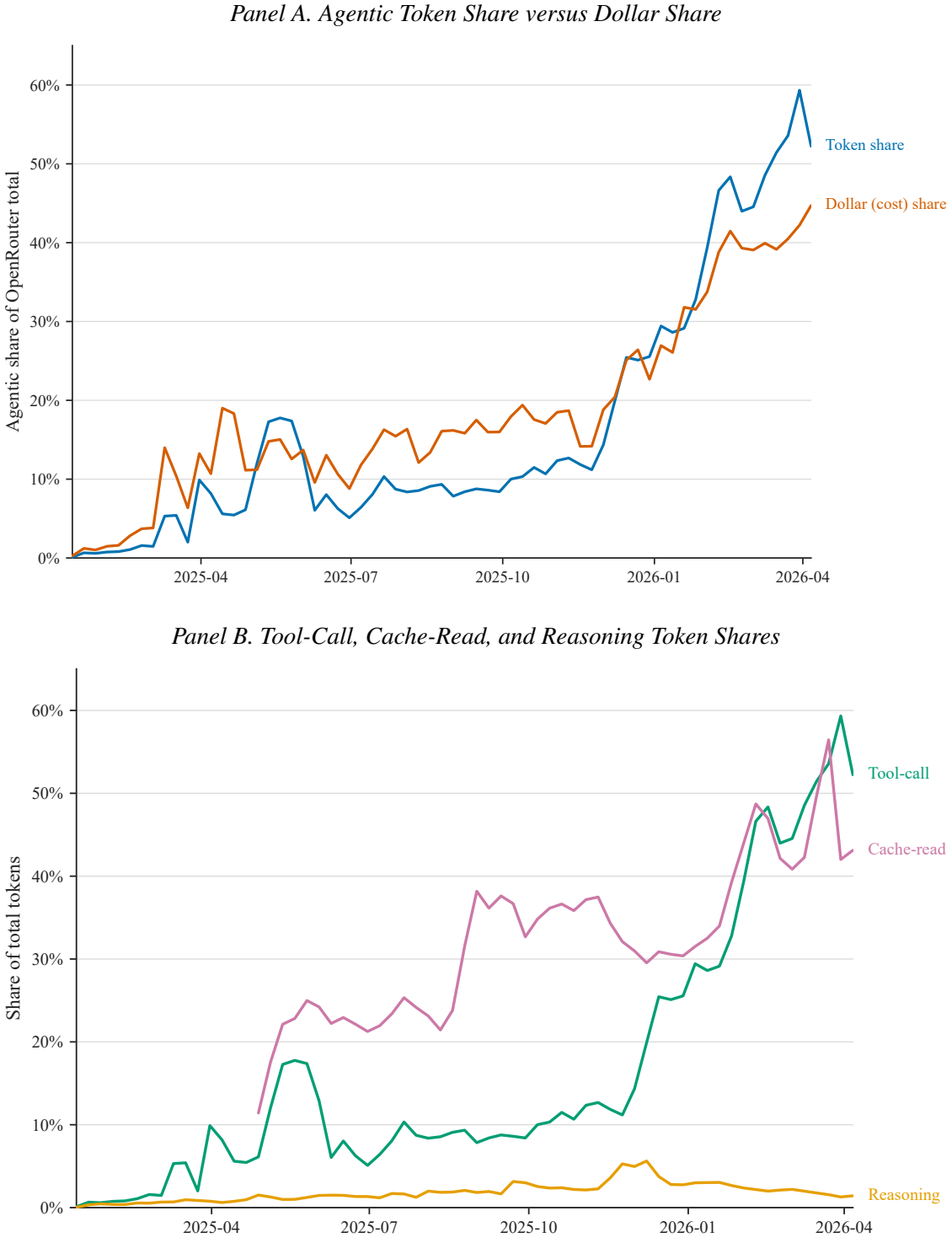


Figure 7: Token, Dollar, and Intensity Shares of Agentic AI Consumption

The two panels plot the weekly evolution of agentic AI consumption in the OpenRouter data, January 2024 through April 2026. Panel A compares the agentic share of total tokens with its share of total usage dollars. The agentic margin is defined as all tokens on requests whose normalized finish reason is `tool_calls`. Panel B plots three token-level margins, each as a share of total tokens: the agentic share, the cache-read share (cached tokens divided by total tokens), and the reasoning share (reasoning tokens divided by total tokens).

References

Acemoglu D, Aghion P, Zilibotti F. 2006. Distance to frontier, selection, and economic growth. *Journal of the European Economic Association* **4**: 37–74.

Acemoglu D, Autor D. 2011. Skills, tasks and technologies: Implications for employment and earnings. In Ashenfelter O, Card D (eds.) *Handbook of Labor Economics*, volume 4B. Amsterdam: Elsevier, 1043–1171.

Acemoglu D, Restrepo P. 2019a. Artificial intelligence, automation, and work. In Agrawal A, Gans J, Goldfarb A (eds.) *The Economics of Artificial Intelligence: An Agenda*, chapter 8. Chicago, IL: University of Chicago Press, 197–236.

Acemoglu D, Restrepo P. 2019b. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives* **33**: 3–30.

Acemoglu D, Restrepo P. 2022. Tasks, automation, and the rise in U.S. wage inequality. *Econometrica* **90**: 1973–2016.

Acemoglu D, Zilibotti F. 1997. Was Prometheus unbound by chance? Risk, diversification, and growth. *Journal of Political Economy* **105**: 709–751.

Aghion P, Howitt P. 1992. A model of growth through creative destruction. *Econometrica* **60**: 323–351.

Andrews I, Farboodi M. 2025. Do markets believe in transformative ai? Technical report, Working paper.

Anthropic. 2026. Anthropic economic index report: Learning curves. `https://www.anthropic.com/research/economic-index-march-2026-report`. Accessed June 2026.

Aubakirova M, Atallah A, Clark C, Summerville J, Midha A. 2025. State of AI: An Empirical 100 Trillion Token Study with OpenRouter. OpenRouter and a16z report, arXiv:2601.10088.

Autor DH, Levy F, Murnane RJ. 2003. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics* **118**: 1279–1333.

Babina T. 2026. Understanding firms' ai efforts and their economic impact. Technical Report 35123, National Bureau of Economic Research.

Babina T, Fedyk A, He A, Hodson J. 2024a. Artificial intelligence and firms' systematic risk. Technical report, Social Science Research Network.

Babina T, Fedyk A, He A, Hodson J. 2024b. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics* **151**: 103745.

Carhart MM. 1997. On persistence in mutual fund performance. *Journal of Finance* **52**: 57–82.

Das D, Murphy M. 2026. OpenRouter Now Processes More Than a Quadrillion Tokens a Year. Menlo Ventures.

Deming DJ. 2017. The growing importance of social skills in the labor market. *Quarterly Journal of Economics* **132**: 1593–1640.

Demirer M, Fradkin A, Tadelis N, Peng S. 2025. The emerging market for intelligence: Pricing, supply, and demand for llms. Technical Report 34608, National Bureau of Economic Research.

Eisfeldt AL, Schubert G, Zhang MB. 2023. Generative ai and firm values. Technical report, National Bureau of Economic Research.

Eloundou T, Manning S, Mishkin P, Rock D. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. Working paper, OpenAI. ArXiv:2303.10130.

Fama EF, French KR. 1992. The cross-section of expected stock returns. *Journal of Finance* **47**: 427–465.

Fama EF, French KR. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**: 3–56.

Fama EF, French KR. 2015. A five-factor asset pricing model. *Journal of Financial Economics* **116**: 1–22.

Fama EF, MacBeth JD. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* **81**: 607–636.

Felten EW, Raj M, Seamans R. 2021. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal* **42**: 2195–2217.

Garleanu N, Kogan L, Panageas S. 2012. Displacement risk and asset returns. *Journal of Financial Economics* **105**: 491–510.

Hassabis D, Kavukcuoglu K. 2024. Introducing gemini 2.0: Our new ai model for the agentic era. Google Blog.

Jegadeesh N, Titman S. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* **48**: 65–91.

Jensen TI, Kelly B, Pedersen LH. 2023. Is there a replication crisis in finance? *Journal of Finance* **78**: 2465–2518.

Jones CI, Tonetti C. 2026. Past automation and future A.I.: How weak links tame the growth explosion. Working paper, Stanford GSB and NBER.

Kogan L, Papanikolaou D. 2014. Growth opportunities, technology shocks, and asset prices. *Journal of Finance* **69**: 675–718.

Kogan L, Papanikolaou D, Schmidt LD, Seegmiller B. 2023. Technology and labor displacement: Evidence from linking patents with worker-level data. Technical report, National Bureau of Economic Research.

Kogan L, Papanikolaou D, Seru A, Stoffman N. 2017. Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics* **132**: 665–712.

Ng A. 2024a. Agentic design patterns part 1: Four ai agent strategies that improve gpt-4 and gpt-3.5 performance. The Batch, DeepLearning.AI.

Ng A. 2024b. Agentic design patterns part 3: Tool use. The Batch, DeepLearning.AI.

Pastor L, Veronesi P. 2009. Technological revolutions and stock prices. *American Economic Review* **99**: 1451–1483.

Pichai S. 2026. Google I/O 2026: Sundar Pichai's opening keynote. `https://blog.google/innovation-and-ai/sundar-pichai-io-2026/`. Accessed June 2026.

Tan H. 2026. Goldman Sachs Says the AI Boom Is Bigger Than Investors Think. Business Insider.

Webb M. 2020. The impact of artificial intelligence on the labor market. Working paper, Stanford University.

Yang J, Yonack N, Zyskowski K, Yarats D, Ho J, Ma J. 2025. The adoption and usage of AI agents: Early evidence from Perplexity. ArXiv:2512.07828.

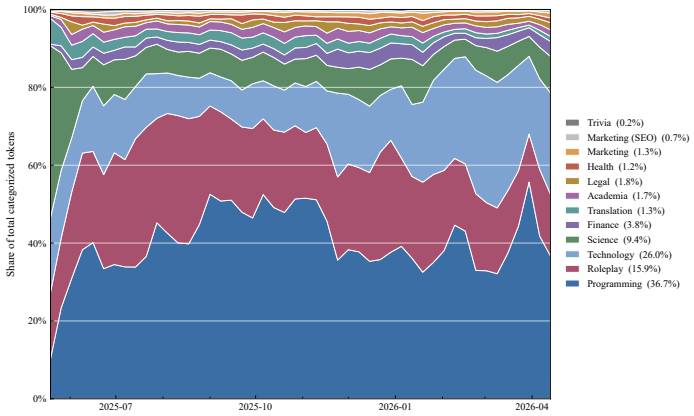
Online Appendix

This is the Online Appendix to Borri, Liu and Tsyvinski, "AI Premium"

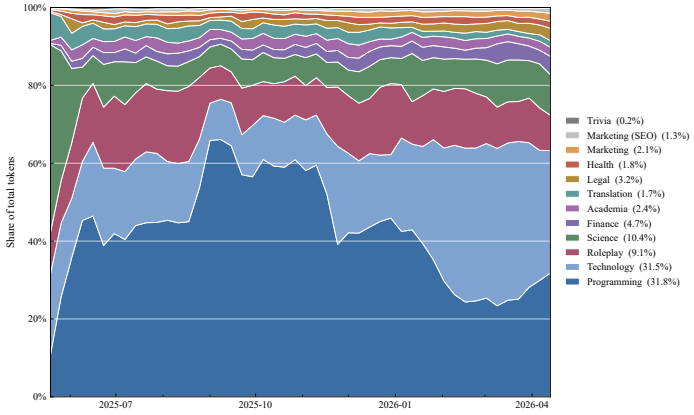
IA.1 AI Data

This section documents the scale and broad composition of the OpenRouter, Inc. data used to construct the AI token consumption measures. Figure 1 shows the rapid growth in total-token consumption on the platform and reports cumulative total-token consumption over the sample. Cumulative tokens rise to 380.8 trillion by the end of the sample, with most of the growth concentrated in the second half of the period and no obvious single-event jump. Figure IA.1 decomposes AI consumption across twelve prompt-content categories, separately for all models, closed-source models, and open-source models. The prompt labels are assigned by OpenRouter as discussed in Aubakirova, Atallah, Clark, Summerville and Midha (2025). Across all models, programming and technology together account for the majority of categorized tokens at the end of the sample (37 percent and 26 percent), followed by roleplay and science. Open-weight models tilt more toward programming and roleplay. Closed-source models tilt more toward technology. Category coverage begins in May 2025, when OpenRouter started populating the field.

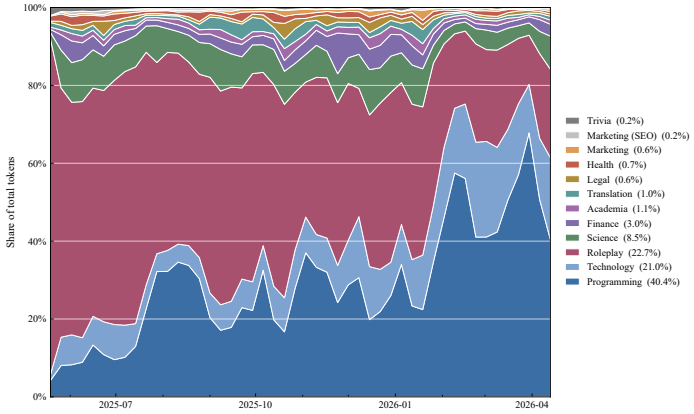
Table IA.1 reports summary statistics for OpenRouter consumption at the account-day, account, and account-model levels, corresponding to the highest level of granularity of the data we use. Panel A summarizes active account-days, Panel B summarizes account-level breadth and concentration, and Panel C summarizes account-model daily consumption and volatility.



(A) All AI Models



(B) Closed-Source Models Only



(C) Open-Source Models Only

Figure IA.1: Evolution of AI Consumption by Prompt-Content Category

Each panel in the figure stacks the weekly share of total categorized tokens contributed by each of the 12 prompt-content categories tagged in the platform's category-token-share file. Panel A pools all AI models; Panel B restricts the universe to closed-weight models; Panel C restricts the universe to open-weight models. Within each panel shares sum to 100% by construction, and the right-hand legend reports the share at the final week (legend ordering matches the stacking order). Data is weekly from mid-May 2025 (when category coverage becomes available) through April 2026.

Table IA.1: OpenRouter Account-Model Data Summary

	Mean	Std.	p10	p25	p50	p75	p90
Panel A. Active account-day consumption							
Requests per active account-day	109.9	434.3	1.000	4.000	16.00	49.00	153.0
Total tokens per active account-day (thous.)	1,223	4,778	2.718	14.55	79.44	392.7	1,713
User-days using at least two models (%)	31.66						
User-days using at least two providers (%)	44.43						
Panel B. Account-level breadth and persistence							
Active days per account	16.73	35.73	1.000	1.000	3.000	13.00	47.00
Active weeks per account	5.300	8.131	1.000	1.000	2.000	5.000	15.00
Observed account span (days)	67.24	112.1	1.000	1.000	8.000	84.00	243.0
Lifetime requests	1,127	4,470	2.000	9.000	56.00	348.0	1,683
Lifetime total tokens (thous.)	15,916	64,301	2.258	22.36	288.7	2,723	22,607
Top-model token share within account (%)	79.40	24.56	40.19	59.35	93.15	100.0	100.0
Model-token HHI within account	7,324	2,945	2,791	4,830	8,712	10,000	10,000
Accounts using at least two models (%)	60.37						
Accounts using at least five models (%)	29.73						
Accounts using at least two providers (%)	70.00						
Panel C. Account-model daily consumption and volatility							
Active days per account-model pair	5.066	10.07	1.000	1.000	1.000	4.000	12.00
Observed pair span (days)	20.79	44.47	1.000	1.000	1.000	16.00	65.00
Active-day share within pair span (%)	74.38	35.73	12.50	41.18	100.0	100.0	100.0
Requests per active account-model-day	24.12	73.73	1.000	1.500	4.000	14.60	44.50
Total tokens per active account-model-day (thous.)	372.9	1,424	0.419	2.721	16.33	109.2	558.9
SD of daily total tokens (thous.)	971.0	3,385	3.401	13.28	67.03	311.4	1,614
CV of daily total tokens	1.132	0.561	0.561	0.749	1.005	1.379	1.865

This table reports account-level and account-model-level summary statistics computed from the OpenRouter generations data. Non-share row variables are winsorized at the 1st and 99th percentiles before computing the reported statistics. An active account-day has positive recorded requests; requests are API calls; total tokens are prompt plus completion tokens. Active days and weeks count positive-use days and calendar weeks, and observed span is first-to-last positive-use days inclusive. Top-model share and model-token HHI use each account's lifetime token shares across models, with HHI scaled by 10,000. For account-model pairs, active-day share is active days divided by span; SD and CV are computed across active days, with CV equal to SD divided by the mean. Volatility rows require at least five active days. Share rows are percent means and leave distribution columns blank.

IA.2 AI Factor

This appendix documents the construction of the AI factor introduced in Section 2.1 and of the split-specific factors used in Section 3.3.3. Figure 1 of the main text plots total tokens in levels; we begin in growth space, which is what enters the construction of the factor.

Figure IA.2 plots the weekly log growth of total tokens, usage dollars, and distinct users after full-sample standardization. The three series move closely together but are not identical, motivating the use of their common principal component.

Table IA.2 then reports the principal-component loadings used to combine those three growth series into the baseline AI factor of Section 2.1. The first principal component places weights of 0.665, 0.559, and 0.496 on total tokens, dollars, and users, and explains 56.5 percent of the joint variance. The roughly equal loadings indicate that a single common factor captures the bulk of co-movement across the three margins.

Table IA.2: PCA Loadings for Weekly Log Growth in AI Variables

	Comp. 1	Comp. 2	Comp. 3
$\Delta \ln Tok$	0.665	−0.047	−0.745
$\Delta \ln Dol$	0.559	−0.631	0.538
$\Delta \ln User$	0.496	0.774	0.394
Eigenvalue	1.695	0.858	0.447
Proportion	*0.565*	*0.286*	*0.149*
Cumulative	*0.565*	*0.851*	*1.000*

The table reports the loadings of each principal component on the three underlying input series $\Delta \ln Tok$, $\Delta \ln Dol$, and $\Delta \ln User$, where the inputs are the log growth in total tokens, usage dollars, and distinct active users, respectively. Panel A is the principal component decomposition at the daily frequency. Panel B is the weekly Monday-indexed decomposition used to construct the headline asset-pricing factor AI_t (the first component of weekly inputs). The *Proportion* and *Cumulative* rows give the share of the input variance captured by each principal component. Loadings are in correlation form (eigenvectors of the input correlation matrix, with the sum of squared loadings within each component equal to 1). Data is weekly from January 2024 through April 2026.

We next turn to the four user- and model-side splits that underlie the split-specific factors used in Table 7 of Section 3.3.3. Figure IA.3 plots each as a weekly token-share time series: the closed-source share, the paid/core account share, the seasoned-user share, and the long-prompt share. The four series are distinct and not collinear—the closed-source share fluctuates with frontier-model release cycles, the paid/core and seasoned-user shares evolve more smoothly, and the long-prompt share is the most volatile—so each generates a different split-specific AI factor in the main analysis.

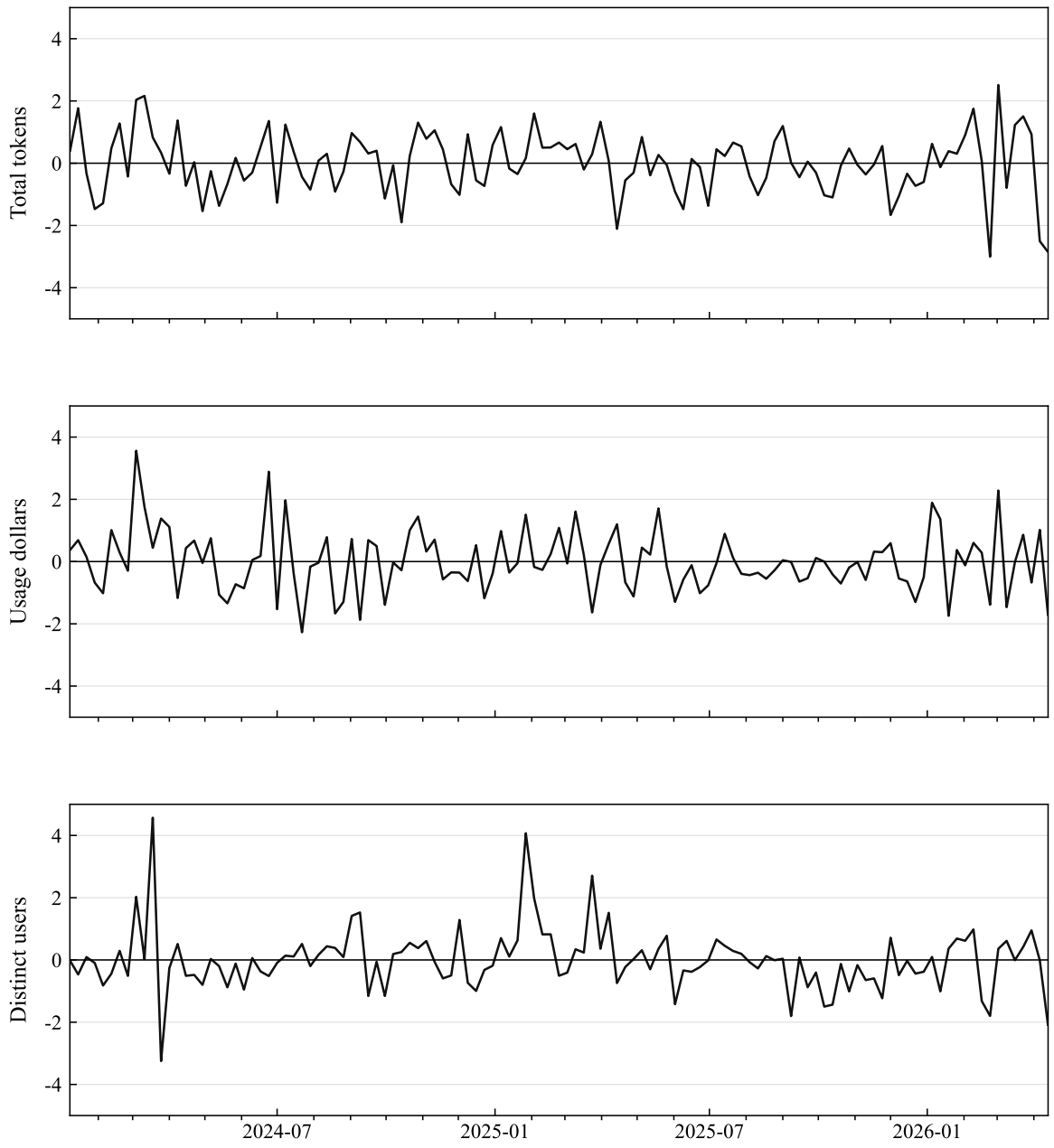
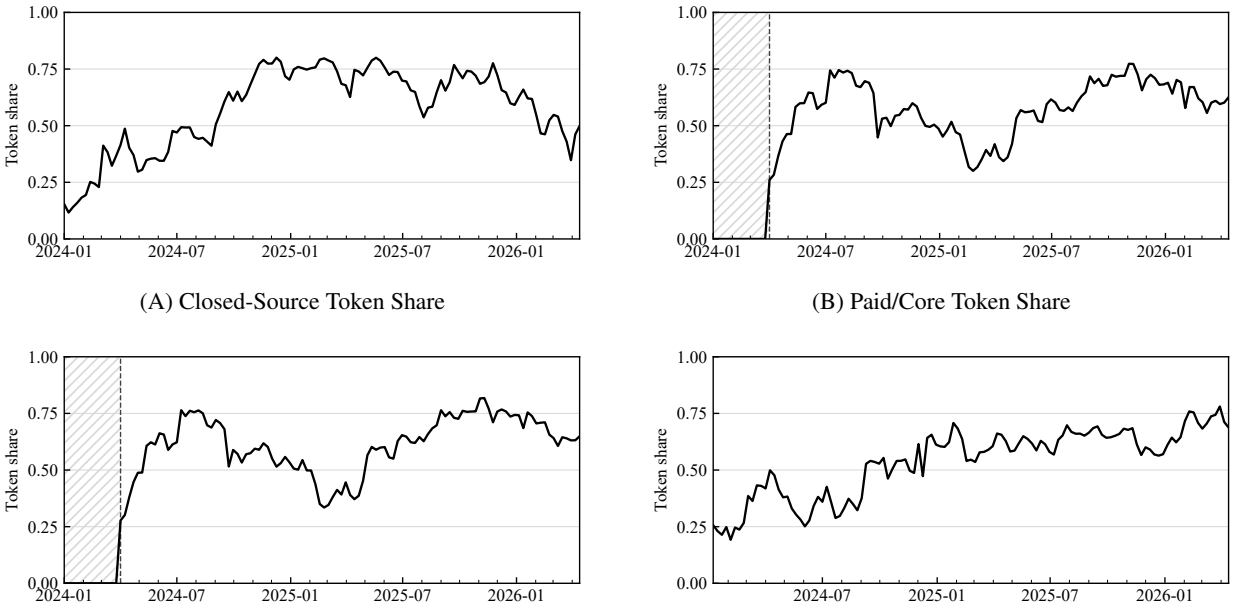


Figure IA.2: Standardized Weekly Natural-Log Growth of the Three AI-Factor Input Series

The figure plots the full-sample-standardized weekly natural-log growth of the three input components used to construct the headline AI factor: total tokens, usage dollars, and distinct users. The corresponding weekly levels are plotted in main-paper Figure 1. Data is weekly from January 2024 through April 2026.



(A) Closed-Source Token Share

(B) Paid/Core Token Share

(C) Seasoned-User Token Share

(D) Long-Prompt (Top 25%) Token Share

Figure IA.3: Evolution of the AI Consumption Decompositions Used in Table 7

Each panel plots, for one of the four splits used in Table 7, the weekly share of total tokens attributable to the focal leg of that split: Panel A, closed-source AI models (Table 7 Panel A); Panel B, paid/core accounts under the `acct_paid` definition (Table 7 Panel B); Panel C, seasoned users (Table 7 Panel C); Panel D, the long-prompt segment defined as rows whose prompt-token count exceeds the prior-13-week rolling 75th percentile (Table 7 Panel D). Hatched regions and dashed vertical lines in Panels B and C mark burn-in periods for the user classifications.

IA.3 AI Portfolios

This section collects portfolio evidence and robustness tests that complement the main-text sorts. We organize the discussion around four questions about the headline AI-premium result: whether it depends on how the AI factor is built, whether it survives standard risk adjustment and small-sample inference, whether it merely relabels broad technology, attention, or disclosure exposures, and how it varies across geography and the rise of agentic AI consumption.

We start with the factor construction. Table IA.3 repeats the value-weighted, all-stock-breakpoint quintile sort using each of the three input series ($\Delta \ln Tok$, $\Delta \ln Dol$, $\Delta \ln User$) in place of the principal-component AI factor. The high-minus-low spreads are 51.6, 36.8, and 46.3 basis points per week, so the headline result is not driven by a single input.

Table IA.3: AI-Factor Input-Component Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Sorted on β^{Tok}						
Excess return	0.042	0.346	0.419	0.270	0.558	0.516
t	(0.16)	(1.50)	(2.02)	(1.24)	(1.79)	(2.35)
FF5 α	-0.157	0.091	0.157	0.028	0.198	0.356
$t(\alpha)$	(-1.20)	(1.06)	(2.26)	(0.35)	(1.42)	(1.64)
FF5+Mom α	-0.152	0.095	0.154	0.027	0.198	0.350
$t(\alpha)$	(-1.19)	(1.16)	(2.26)	(0.33)	(1.41)	(1.63)
Panel B. Sorted on β^{Dol}						
Excess return	0.213	0.258	0.377	0.416	0.580	0.368
t	(0.81)	(1.13)	(1.83)	(1.78)	(1.91)	(1.43)
FF5 α	0.003	0.032	0.144	0.150	0.288	0.284
$t(\alpha)$	(0.02)	(0.41)	(2.05)	(1.47)	(1.98)	(1.13)
FF5+Mom α	0.000	0.034	0.146	0.149	0.287	0.287
$t(\alpha)$	(0.00)	(0.44)	(2.08)	(1.45)	(1.97)	(1.13)
Panel C. Sorted on β^{User}						
Excess return	0.137	0.439	0.233	0.375	0.600	0.463
t	(0.49)	(1.83)	(1.18)	(1.76)	(1.89)	(2.21)
FF5 α	-0.124	0.164	-0.002	0.144	0.298	0.422
$t(\alpha)$	(-0.95)	(1.93)	(-0.03)	(1.53)	(2.05)	(1.95)
FF5+Mom α	-0.117	0.165	0.000	0.144	0.294	0.412
$t(\alpha)$	(-0.93)	(1.93)	(0.00)	(1.52)	(2.03)	(1.96)

The table decomposes the AI-factor result of Table 3 into its three input components. The AI factor is the first principal component of weekly log growth in total tokens, usage dollars, and distinct users. Here we replace the AI factor with each component individually: for $X \in \{Tok, Dol, User\}$, we estimate $r^e_{i,t} = a + b_{\text{mkt}} r^e_{m,t} + \beta^X_{i,t} \Delta \ln X_t$ on a 13-week rolling window with at least 9 non-missing weeks, then sort stocks into pooled quintiles by the estimated $\beta^X_{i,t}$. Returns and alphas are in percent per week, value-weighted by formation-week market capitalization, with all-stock breakpoints; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and from that model augmented with Carhart, 1997 momentum. Data is weekly from January 2024 through April 2026.

Turning from factor construction to risk adjustment, Table IA.4 reports the time-series regression of the value-weighted, all-stock-breakpoint AI-beta long-short return on the FF3, FF5, and FF5-plus-momentum models. The intercept is 58.2, 56.3, and 55.9 basis points per week respectively, all significant at the 5

percent level, and none of the factor loadings is large or statistically significant. The H−L spread is not absorbed by size, value, profitability, investment, or momentum.

Table IA.4: Fama-French Factor Regressions for the AI-Beta H−L Portfolio

	FF3	FF5	FF5+Mom
α	0.582	0.563	0.559
$t(\alpha)$	(2.55)	(2.43)	(2.41)
Mkt-RF	0.184	0.216	0.195
t	(1.64)	(1.79)	(1.59)
SMB	0.021	0.072	0.116
t	(0.13)	(0.39)	(0.61)
HML	0.252	0.270	0.273
t	(1.50)	(1.51)	(1.53)
RMW		0.088	0.139
t		(0.37)	(0.57)
CMA		-0.205	-0.122
t		(-0.88)	(-0.49)
Mom			0.146
t			(1.00)
R^2	0.052	0.061	0.072

The dependent variable is the value-weighted all-stock-breakpoint H−L return from Table 3. Returns and factors are in percent per week. FF3 is the Fama–French three-factor model (Fama and French, 1993). FF5 adds RMW and CMA (Fama and French, 2015). FF5+Mom additionally includes the momentum factor (Carhart, 1997). The table reports OLS coefficient estimates, t-statistics in parentheses and R^2. Data is weekly from January 2024 through April 2026.

Beyond the headline portfolio sorts, Table IA.5 looks directly at incumbent application-layer firms. It reports value-weighted SaaS-portfolio regressions on the baseline AI factor and the market. Over the full sample the SaaS-portfolio loading on AI_t is 0.035 ($t = 0.27$), but restricted to the last calendar quarter the loading flips sign to -1.353 ($t = -2.69$) per one-standard-deviation AI shock. Public SaaS valuations have shifted from a small positive to a notable negative covariance with realized AI consumption growth, consistent with markets pricing AI-driven displacement risk for some incumbent application-layer firms.

Table IA.5: SaaS Exposure to the AI Factor: Full Sample Versus Last Quarter

	Full sample	Last quarter
Value-weighted SaaS portfolio regression		
AI_t	0.035	-1.353
t	(0.27)	(-2.69)
Mkt-RF	1.137	1.435
t	(14.85)	(4.35)
R^2	0.698	0.714
Avg. firms/week	140.9	142.0

The table reports value-weighted portfolio regressions for public SaaS firms. The SaaS universe is the SaaSDB public-company list matched by ticker to the CRSP/firm-exposure sample. The dependent variable is the value-weighted SaaS portfolio excess log return, and the regressors are the baseline AI factor AI_t and the weekly market excess return Mkt-RF. Standard errors use Newey-West four-week lags. The coefficient on AI_t is multiplied by $100 \times \sigma_w(AI)$, so the row reports the weekly return effect in percentage points for a one-standard-deviation increase in the baseline AI factor. Data is weekly from January 2024 through April 2026.

We next address two inference-related concerns about the headline test. Table IA.6, Panel A, reports

a circular moving-block bootstrap that resamples consecutive holding-week observations: the [2.5, 97.5]-percentile interval for the H−L excess return is [27.3, 101.9] basis points per week, with similar coverage for the FF5 and FF5+Mom alphas, so the interval excludes zero in all three cases. Panel B replaces the raw rolling AI beta with an empirical-Bayes-shrunk beta toward the cross-sectional mean. The value-weighted, all-stock-breakpoint H−L spread is 48.2 basis points per week ($t = 2.42$), so the result is not an artifact of estimation noise in the firm-level betas.

Table IA.6: Small-Sample Robustness of AI-Beta Portfolio

Panel A. Baseline H−L: Block-Bootstrap Inference				
	Estimate	2.5%	97.5%	t
Excess return	0.641	0.273	1.019	(2.84)
FF5 α	0.563	0.128	0.987	(2.43)
FF5+Mom α	0.559	0.118	1.018	(2.41)
Panel B. H−L Sorts on Shrunk AI Beta				
	VW All	VW NYSE	EW All	EW NYSE
Excess return	0.482	0.422	0.263	0.270
t	(2.42)	(2.23)	(1.65)	(1.68)
FF5 α	0.437	0.371	0.199	0.209
$t(\alpha)$	(2.14)	(1.93)	(1.26)	(1.31)
FF5+Mom α	0.430	0.366	0.192	0.202
$t(\alpha)$	(2.13)	(1.91)	(1.25)	(1.30)

The table examines the robustness of the baseline H−L portfolio in Table 3. Panel A reports percentile confidence intervals from a circular moving-block bootstrap that resamples consecutive holding-week observations. Panel B repeats the quintile sort after replacing the raw rolling AI beta with an empirical-Bayes-shrunk beta. The shrinkage target is the formation-week cross-sectional mean beta, and the reliability weight uses the rolling-beta standard error and the formation-week cross-sectional signal variance. Returns and alphas are in percent per week; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and from that model augmented with Carhart, 1997 momentum. Breakpoint and weighting definitions match Table 3. Data is weekly from January 2024 through April 2026.

We then ask whether the AI premium is simply a relabeling of broader cross-sectional patterns. Industry tilts are the first candidate. Table IA.7 replaces the AI beta with its deviation from the within-period Fama-French 30 industry median and re-runs the quintile sort. The value-weighted, all-stock-breakpoint H−L spread is 60.3 basis points per week ($t = 3.03$), close to the 64.1-basis-point baseline, so the AI premium is a within-industry phenomenon and not a relabeling of broad industry tilts.

Broad public attention to AI is the second candidate. Table IA.8 re-estimates each firm's AI beta after adding the weekly log growth of Google Trends search interest in "Artificial intelligence" to the first-stage market regression, and sorts on the residual AI beta. The value-weighted, all-stock-breakpoint H−L spread is 69.6 basis points per week ($t = 3.35$), with comparable factor-model alphas, so the premium is not a relabeling of broad public attention to AI.

A third candidate is firms' own AI narratives. Table IA.9 regresses the baseline AI-beta H−L return on a control H−L portfolio built from firms' AI mentions in 10-K, 10-Q, and 8-K filings, using unscaled AI-mention counts in Panel A and word-scaled counts in Panel B. The spanning intercept ranges from 56.5 to 66.8 basis points per week, comparable to the unconditional H−L return, and the SEC-mention loading is small and statistically insignificant in all rows. The realized-consumption signal is distinct from firms' own AI narratives in their disclosures.

Table IA.7: Industry-Demeaned AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Value-weighted, all-stock breakpoints						
Excess return	0.056	0.334	0.254	0.427	0.659	0.603
t	(0.20)	(1.58)	(1.17)	(1.78)	(2.21)	(3.03)
FF5 α	-0.206	0.106	0.012	0.181	0.343	0.550
$t(\alpha)$	(-1.56)	(1.31)	(0.16)	(1.96)	(2.92)	(2.67)
FF5+Mom α	-0.200	0.107	0.012	0.181	0.341	0.542
$t(\alpha)$	(-1.56)	(1.32)	(0.17)	(1.94)	(2.90)	(2.68)
Panel B. Value-weighted, NYSE breakpoints						
Excess return	0.050	0.331	0.223	0.417	0.701	0.652
t	(0.19)	(1.56)	(1.01)	(1.83)	(2.44)	(3.61)
FF5 α	-0.189	0.093	-0.019	0.170	0.394	0.583
$t(\alpha)$	(-1.69)	(1.17)	(-0.24)	(2.01)	(3.59)	(3.15)
FF5+Mom α	-0.185	0.094	-0.018	0.170	0.392	0.577
$t(\alpha)$	(-1.69)	(1.19)	(-0.22)	(1.99)	(3.57)	(3.16)
Panel C. Equal-weighted, all-stock breakpoints						
Excess return	0.136	0.270	0.251	0.312	0.403	0.267
t	(0.43)	(1.11)	(1.06)	(1.21)	(1.30)	(1.89)
FF5 α	-0.028	0.114	0.090	0.122	0.192	0.220
$t(\alpha)$	(-0.31)	(2.54)	(2.43)	(2.57)	(2.12)	(1.51)
FF5+Mom α	-0.023	0.116	0.090	0.122	0.189	0.212
$t(\alpha)$	(-0.27)	(2.59)	(2.45)	(2.56)	(2.11)	(1.52)
Panel D. Equal-weighted, NYSE breakpoints						
Excess return	0.142	0.270	0.274	0.300	0.405	0.263
t	(0.46)	(1.13)	(1.15)	(1.19)	(1.35)	(2.01)
FF5 α	-0.022	0.121	0.109	0.115	0.197	0.219
$t(\alpha)$	(-0.28)	(2.75)	(2.81)	(2.40)	(2.36)	(1.63)
FF5+Mom α	-0.018	0.122	0.110	0.115	0.195	0.213
$t(\alpha)$	(-0.23)	(2.80)	(2.83)	(2.39)	(2.35)	(1.64)

The table repeats the headline test of Table 3 after demeaning the sorting variable by FF30 industry. For each formation week t and FF30 industry j we compute the cross-sectional median AI beta $\widetilde{\beta}_t^{AI,(j)} = \text{median}_{i \in j}(\beta_{i,t}^{AI})$ across all firms in industry j that week, requiring at least five eligible firms per industry-week. The industry-demeaned beta is $\beta_{i,t}^{AI,Ind} = \beta_{i,t}^{AI} - \widetilde{\beta}_t^{AI,(j(i))}$, and stocks are sorted into quintiles on this within-industry deviation. Industry assignment follows the Kenneth French convention: the SIC code from the most recent Compustat fiscal year ending in calendar year $t-1$ (CRSP June-of-t SICCD as fallback) is mapped to one of the 30 Fama-French industries via French's `Siccodes30`. Returns and alphas are in percent per week; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and from that model augmented with Carhart, 1997 momentum. Data is weekly from January 2024 through April 2026.

Stepping back to factor construction, Table IA.10 replaces the principal-component AI factor with an alternative aggregator: the simple equal-weighted average of the three full-sample-standardized input series. The value-weighted, all-stock-breakpoint H−L spread is 64.9 basis points per week ($t = 2.78$), with FF5 and FF5+Mom alphas of 57.6 and 57.2 basis points respectively. The headline result is not specific to the principal-component construction.

We also split AI consumption into agentic and non-agentic components and sort on exposure to each separately. Table IA.11 builds two alternative factors from token growth split this way: agentic tokens come from requests with normalized finish reason `tool_calls`, in which the model invoked an external tool,

Table IA.8: Google-Trends-Controlled AI-Beta Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Value-weighted, all-stock breakpoints						
Excess return	0.025	0.264	0.283	0.482	0.721	0.696
t	(0.09)	(1.16)	(1.31)	(2.11)	(2.46)	(3.35)
FF5 α	-0.221	0.036	0.033	0.225	0.397	0.618
$t(\alpha)$	(-1.67)	(0.39)	(0.45)	(2.35)	(3.00)	(2.87)
FF5+Mom α	-0.216	0.040	0.033	0.224	0.393	0.609
$t(\alpha)$	(-1.66)	(0.43)	(0.44)	(2.33)	(3.00)	(2.89)
Panel B. Value-weighted, NYSE breakpoints						
Excess return	0.121	0.233	0.234	0.565	0.709	0.588
t	(0.48)	(1.04)	(1.08)	(2.51)	(2.47)	(2.98)
FF5 α	-0.101	-0.006	-0.012	0.305	0.391	0.491
$t(\alpha)$	(-0.86)	(-0.07)	(-0.16)	(3.29)	(3.11)	(2.45)
FF5+Mom α	-0.096	-0.003	-0.013	0.304	0.388	0.484
$t(\alpha)$	(-0.83)	(-0.04)	(-0.18)	(3.26)	(3.09)	(2.46)
Panel C. Equal-weighted, all-stock breakpoints						
Excess return	-0.032	0.225	0.265	0.279	0.299	0.331
t	(-0.10)	(0.94)	(1.25)	(1.14)	(0.97)	(2.08)
FF5 α	-0.184	0.069	0.116	0.090	0.093	0.277
$t(\alpha)$	(-1.76)	(1.35)	(2.82)	(1.67)	(0.93)	(1.72)
FF5+Mom α	-0.180	0.071	0.117	0.090	0.091	0.272
$t(\alpha)$	(-1.76)	(1.44)	(2.83)	(1.65)	(0.91)	(1.71)
Panel D. Equal-weighted, NYSE breakpoints						
Excess return	0.021	0.219	0.275	0.281	0.293	0.272
t	(0.07)	(0.94)	(1.29)	(1.17)	(0.97)	(1.92)
FF5 α	-0.131	0.069	0.128	0.095	0.088	0.219
$t(\alpha)$	(-1.49)	(1.41)	(3.21)	(1.83)	(0.96)	(1.52)
FF5+Mom α	-0.128	0.071	0.129	0.095	0.086	0.214
$t(\alpha)$	(-1.48)	(1.48)	(3.22)	(1.82)	(0.94)	(1.51)

The table extends the analysis of the properties of the AI-Beta quintile portfolios presented in Table 3 after expanding the first-pass rolling beta regression to include the weekly log growth in Google Trends search interest for the topic *Artificial intelligence* as an additional control: $r^e_{i,t} = a + b_{\text{mkt}}\, r^e_{m,t} + b_{\text{gtrend}}\, \Delta \ln \text{gtrend}^{\text{AI}}_t + b_{\text{AI}}\, AI_t$, with at least 9 non-missing weeks. Stocks are sorted into pooled quintiles by the Google-Trends-controlled b_{AI}. Returns and alphas are in percent per week; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and from that model augmented with Carhart, 1997 momentum. The Google Trends weekly index covers 2023-10 to 2026-02 and is fetched via the Knowledge-Graph topic API. Data is weekly from January 2024 through April 2026.

and non-agentic tokens are the remainder. Sorting on exposure to non-agentic token growth yields an H−L spread of 44.6 basis points per week ($t = 2.15$). Sorting on exposure to agentic token growth yields a smaller and statistically insignificant 31.0 basis points ($t = 0.76$). Agentic use is a fast-growing but still small share of total consumption in this sample, and the priced AI signal loads on broader, non-agentic activity.

Returning to the headline sort, Table IA.12 replaces the quintile portfolios with deciles. The value-weighted, all-stock-breakpoint H−L spread widens to 97.3 basis points per week ($t = 3.31$), with an FF5+Mom alpha of 82.9 basis points ($t = 2.85$). The relationship between AI exposure and future returns is approximately monotonic and strengthens at the extreme tails.

Table IA.9: Spanning Regressions with SEC-Based AI Mentions

	Model	α	$t(\alpha)$	b_{Mention}	$t(b_{\text{Mention}})$	R^2
Panel A. Controlling for unscaled SEC-based AI mentions						
(1)	Mean	0.645	(2.85)			
(2)	SEC-Based AI Mention	0.668	(2.91)	-0.070	(-0.66)	0.005
(3)	FF5	0.569	(2.45)			0.062
(4)	FF5 + SEC-Based AI Mention	0.574	(2.43)	-0.028	(-0.14)	0.062
(5)	FF5 + Mom	0.565	(2.43)			0.073
(6)	FF5 + Mom + SEC-Based AI Mention	0.577	(2.44)	-0.063	(-0.31)	0.074
Panel B. Controlling for word-scaled SEC-based AI mentions						
(1)	Mean	0.645	(2.85)			
(2)	SEC-Based AI Mention	0.650	(2.85)	-0.031	(-0.25)	0.001
(3)	FF5	0.569	(2.45)			0.062
(4)	FF5 + SEC-Based AI Mention	0.568	(2.44)	0.120	(0.48)	0.064
(5)	FF5 + Mom	0.565	(2.43)			0.073
(6)	FF5 + Mom + SEC-Based AI Mention	0.565	(2.42)	0.052	(0.20)	0.073

The table reports time-series spanning regressions of the baseline AI-beta H−L portfolio return on SEC-based AI-mention H−L portfolios. The dependent variable is the value-weighted H−L return from sorting stocks on the baseline rolling AI beta used in Table 3, using all-stock breakpoints. The SEC-based AI-mention controls are constructed from the completed SEC filing text archive. We map firms to SEC CIKs using the latest non-missing CIK from Compustat quarterly data and use 10-K, 10-Q, and 8-K filing text. AI mentions are exact matches to *artificial intelligence*, *generative AI*, *machine learning*, *deep learning*, *large language model*, *large language models*, *LLM*, *OpenAI*, and *AI model*. For firm i in formation week t, let $H_{i,t}$ be the number of AI phrase mentions over weeks $t-13$ through $t-1$, let $F_{i,t}$ be the number of filings over the same lagged window, and let $W_{i,t}$ be the total number of words in those filings. Panel A sorts on the unscaled measure $\ln(1+H_{i,t})-\ln(1+F_{i,t})$. Panel B sorts on the word-scaled measure $10{,}000\times H_{i,t}/W_{i,t}$. In both panels, the SEC-based AI-mention H−L control buys the high-mention quintile and shorts the low-mention quintile among firms with at least one filing in the lagged window. The reported regression is $R_t^{AI,H-L}=\alpha+b_{\text{Mention}}R_t^{SECMention,H-L}+\delta' X_t+\epsilon_t$, where X_t is empty in row (2), contains the Fama-French five factors (Fama and French, 2015) in row (4), and contains the five factors plus Carhart, 1997 momentum in row (6). Returns and factors are in percent per week. Data is weekly from January 2024 through April 2026.

Table IA.10: Average-Standardized Usage-Growth Quintile Portfolios

	Low	2	3	4	High	H−L
Panel A. Value-weighted, all-stock breakpoints						
Excess return	0.007	0.302	0.277	0.530	0.656	0.649
t	(0.03)	(1.35)	(1.27)	(2.35)	(2.21)	(2.78)
FF5 α	-0.215	0.075	0.012	0.275	0.361	0.576
$t(\alpha)$	(-1.51)	(0.81)	(0.16)	(3.22)	(2.61)	(2.42)
FF5+Mom α	-0.212	0.078	0.012	0.273	0.360	0.572
$t(\alpha)$	(-1.49)	(0.86)	(0.16)	(3.20)	(2.59)	(2.40)
Panel B. Value-weighted, NYSE breakpoints						
Excess return	0.147	0.191	0.368	0.483	0.667	0.520
t	(0.56)	(0.86)	(1.68)	(2.10)	(2.35)	(2.47)
FF5 α	-0.086	-0.020	0.111	0.218	0.372	0.459
$t(\alpha)$	(-0.65)	(-0.21)	(1.43)	(2.49)	(2.89)	(2.13)
FF5+Mom α	-0.082	-0.018	0.111	0.216	0.372	0.455
$t(\alpha)$	(-0.63)	(-0.19)	(1.42)	(2.48)	(2.87)	(2.11)
Panel C. Equal-weighted, all-stock breakpoints						
Excess return	-0.030	0.213	0.267	0.268	0.316	0.345
t	(-0.09)	(0.89)	(1.26)	(1.10)	(1.03)	(2.10)
FF5 α	-0.181	0.059	0.120	0.074	0.112	0.293
$t(\alpha)$	(-1.70)	(1.09)	(2.91)	(1.38)	(1.13)	(1.76)
FF5+Mom α	-0.177	0.062	0.120	0.073	0.111	0.287
$t(\alpha)$	(-1.69)	(1.21)	(2.90)	(1.37)	(1.11)	(1.75)
Panel D. Equal-weighted, NYSE breakpoints						
Excess return	0.028	0.205	0.279	0.264	0.323	0.295
t	(0.09)	(0.88)	(1.29)	(1.09)	(1.08)	(1.96)
FF5 α	-0.124	0.055	0.133	0.073	0.120	0.244
$t(\alpha)$	(-1.29)	(1.12)	(3.17)	(1.37)	(1.32)	(1.59)
FF5+Mom α	-0.120	0.057	0.133	0.073	0.118	0.238
$t(\alpha)$	(-1.28)	(1.20)	(3.15)	(1.35)	(1.30)	(1.58)

The table presents robustness evidence on the properties of the AI-Beta quintile portfolios of Table 3 after replacing the AI factor with a simple equal-weighted average of the three full-sample-standardized input series: $\Delta\widetilde{\ln} X_t = 1/3\left[z(\Delta \ln Tok_t) + z(\Delta \ln Dol_t) + z(\Delta \ln User_t)\right]$, where $z(\cdot)$ denotes full-sample standardization (subtract sample mean, divide by sample standard deviation). This construction partly addresses the forward-looking concern that principal-component weights are estimated from the full sample, although it does not eliminate full-sample information because the input series are still standardized using their full-sample means and standard deviations. The first-pass rolling regression is $r^e_{i,t} = a + b_{\text{mkt}} r^e_{m,t} + b_{\text{avg}} \Delta\widetilde{\ln} X_t$ on a 13-week rolling window with at least 9 non-missing weeks. Stocks are sorted into quintiles by b_{avg}. The investment universe, weighting, and breakpoint definitions match Table 3. Returns and alphas are in percent per week. Data is weekly from January 2024 through April 2026.

Table IA.11: Agentic and Non-Agentic Token-Growth Quintile Portfolios

	Low	2	3	4	High	H–L
Value-weighted, all-stock breakpoints						
$\beta^{Tok,Agentic}$						
Excess return	0.476	0.305	0.427	0.299	0.786	0.310
t	(1.16)	(0.93)	(1.21)	(0.74)	(1.35)	(0.76)
FF5 α	-0.055	-0.062	0.034	-0.169	0.241	0.295
$t(\alpha)$	(-0.27)	(-0.53)	(0.26)	(-1.72)	(1.44)	(0.99)
FF5+Mom α	-0.026	-0.070	0.023	-0.174	0.240	0.266
$t(\alpha)$	(-0.13)	(-0.60)	(0.18)	(-1.74)	(1.41)	(0.90)
$\beta^{Tok,Nonagentic}$						
Excess return	0.075	0.408	0.321	0.339	0.521	0.446
t	(0.28)	(1.77)	(1.57)	(1.53)	(1.73)	(2.15)
FF5 α	-0.132	0.158	0.060	0.085	0.177	0.309
$t(\alpha)$	(-1.07)	(1.77)	(0.89)	(1.08)	(1.35)	(1.50)
FF5+Mom α	-0.128	0.163	0.060	0.084	0.175	0.302
$t(\alpha)$	(-1.05)	(1.92)	(0.88)	(1.06)	(1.34)	(1.49)

The table repeats the value-weighted, all-stock-breakpoint portfolio test of Table 3 after replacing the baseline AI factor with weekly log growth in agentic or non-agentic total tokens. Total tokens are prompt tokens plus completion tokens. Agentic total tokens are total tokens from requests whose normalized finish reason is `tool_calls`; non-agentic total tokens are the remaining total tokens. Firm betas are estimated from a 13-week rolling regression of weekly log-excess returns on the log-excess market return and the token-growth factor, with a minimum of 9 non-missing weeks. Stocks are sorted into quintiles by this beta. Returns and alphas are in percent per week; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and from that model augmented with Carhart, 1997 momentum. Data is weekly from January 2024 through April 2026.

Table IA.12: AI-Beta Decile Portfolios

	Low	2	3	4	5	6	7	8	9	High	H−L
Panel A. Value-weighted, all-stock breakpoints											
Excess return	0.005	0.047	0.420	0.185	0.294	0.259	0.491	0.497	0.381	0.978	0.973
t	(0.01)	(0.18)	(1.84)	(0.75)	(1.33)	(1.09)	(2.07)	(2.09)	(1.40)	(2.63)	(3.31)
FF5 α	-0.251	-0.174	0.175	-0.009	0.088	-0.047	0.241	0.228	0.118	0.591	0.843
$t(\alpha)$	(-1.15)	(-1.15)	(1.43)	(-0.07)	(0.80)	(-0.50)	(2.09)	(2.09)	(0.80)	(2.99)	(2.81)
FF5+Mom α	-0.244	-0.172	0.182	-0.008	0.088	-0.046	0.239	0.226	0.122	0.584	0.829
$t(\alpha)$	(-1.13)	(-1.14)	(1.55)	(-0.06)	(0.80)	(-0.50)	(2.07)	(2.08)	(0.83)	(3.00)	(2.85)
Panel B. Value-weighted, NYSE breakpoints											
Excess return	-0.020	0.260	0.177	0.297	0.374	0.350	0.456	0.462	0.406	0.851	0.871
t	(-0.07)	(1.01)	(0.78)	(1.24)	(1.79)	(1.48)	(2.02)	(1.79)	(1.61)	(2.34)	(3.15)
FF5 α	-0.241	0.020	-0.020	0.107	0.131	0.058	0.193	0.198	0.183	0.471	0.712
$t(\alpha)$	(-1.38)	(0.13)	(-0.17)	(0.86)	(1.31)	(0.55)	(1.90)	(1.51)	(1.38)	(2.41)	(2.59)
FF5+Mom α	-0.237	0.025	-0.017	0.108	0.131	0.058	0.191	0.194	0.186	0.466	0.703
$t(\alpha)$	(-1.36)	(0.17)	(-0.15)	(0.86)	(1.30)	(0.55)	(1.88)	(1.50)	(1.40)	(2.39)	(2.58)
Panel C. Equal-weighted, all-stock breakpoints											
Excess return	-0.166	0.108	0.172	0.207	0.289	0.299	0.228	0.289	0.303	0.340	0.506
t	(-0.45)	(0.39)	(0.70)	(0.89)	(1.32)	(1.41)	(0.99)	(1.11)	(1.06)	(0.99)	(2.32)
FF5 α	-0.302	-0.058	0.008	0.054	0.157	0.156	0.038	0.089	0.100	0.126	0.428
$t(\alpha)$	(-1.99)	(-0.70)	(0.12)	(0.97)	(3.12)	(3.20)	(0.65)	(1.47)	(1.31)	(0.87)	(1.94)
FF5+Mom α	-0.298	-0.054	0.012	0.056	0.158	0.156	0.037	0.088	0.100	0.123	0.421
$t(\alpha)$	(-1.98)	(-0.67)	(0.20)	(1.02)	(3.12)	(3.18)	(0.63)	(1.45)	(1.30)	(0.85)	(1.93)
Panel D. Equal-weighted, NYSE breakpoints											
Excess return	-0.095	0.171	0.232	0.209	0.288	0.304	0.249	0.270	0.301	0.332	0.427
t	(-0.28)	(0.65)	(0.95)	(0.94)	(1.32)	(1.38)	(1.10)	(1.04)	(1.10)	(1.01)	(2.29)
FF5 α	-0.245	0.015	0.075	0.054	0.154	0.156	0.076	0.074	0.090	0.126	0.371
$t(\alpha)$	(-2.02)	(0.18)	(1.33)	(1.08)	(3.05)	(2.87)	(1.29)	(1.20)	(1.36)	(1.00)	(1.95)
FF5+Mom α	-0.241	0.019	0.078	0.057	0.155	0.155	0.075	0.074	0.090	0.124	0.364
$t(\alpha)$	(-2.01)	(0.26)	(1.41)	(1.16)	(3.05)	(2.85)	(1.27)	(1.19)	(1.35)	(0.98)	(1.94)

The table repeats Table 3 with decile, rather than quintile, sorts on 13-week rolling AI beta. Universe, weighting, and breakpoints match Table 3. Returns and alphas are in percent per week; FF5 and FF5+Mom denote alphas from Fama and French, 2015 and that model augmented with Carhart, 1997 momentum. H−L buys High and shorts Low. Data is weekly from January 2024 through April 2026.

A final candidate relabeling is exposure to traded technology baskets. Table IA.13 reports spanning regressions of the AI-beta H−L portfolio return on three benchmarks: the Fama-French 10 HiTec industry return (Panel A), the value-weighted NAICS semiconductor portfolio (Panel B), and an equal-weighted AI/robotics ETF basket (Panel C). Each panel runs the benchmark alone, on top of the market, on top of FF5, and on top of FF5 plus momentum. The spanning intercepts range from 46.1 to 63.5 basis points per week and remain individually significant, and the loadings on the additional benchmarks are small and insignificant in most rows. The AI premium is not absorbed by exposure to technology-sector, semiconductor, or AI-themed equity baskets.

We then turn to geography. Table IA.14 repeats the quintile sort on the Datastream China A-share panel, using either the baseline AI factor (Panel A) or a China-specific AI factor built from OpenRouter consumption originating in mainland China (Panel B). The H−L spreads are uniformly negative and statistically insignificant in both panels and under both weighting schemes. The AI premium is concentrated in developed equity markets near the global AI frontier and is absent from a market where AI is salient to investors but listed firms sit further from that frontier.

Finally, Figure IA.4 plots the 13-week rolling realized volatility of the value-weighted, all-stock-breakpoint AI-beta long-short return, annualized by $\sqrt{52}$. The series varies meaningfully over the sample, with peaks around major AI-news episodes, summarizing how the conditional risk in AI exposure has evolved through the sample.

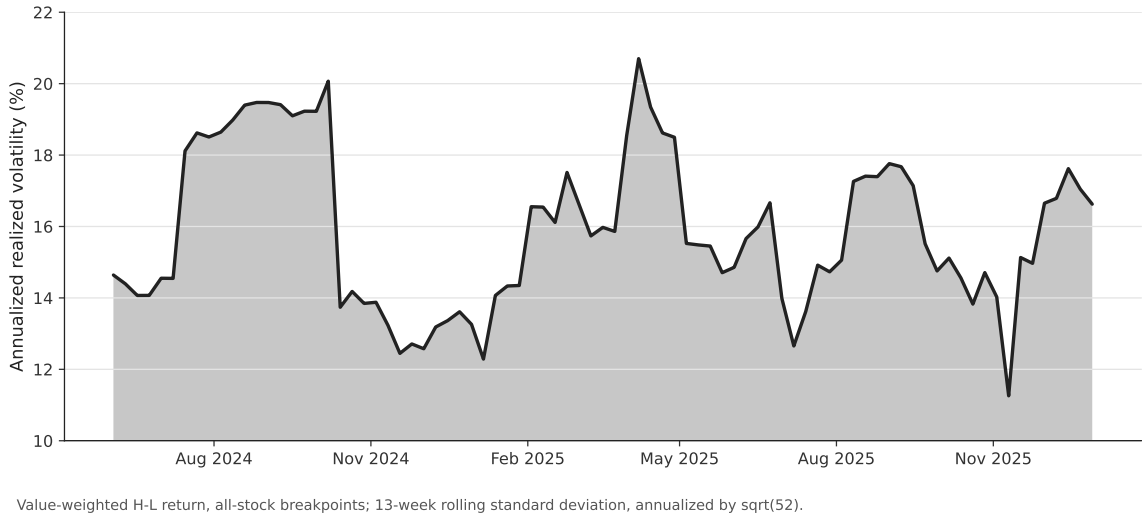


Figure IA.4: AI VIX: Rolling Volatility of the AI-Premium Portfolio

The figure plots the AI VIX index, constructed as the 13-week rolling standard deviation of the value-weighted high-minus-low AI-beta portfolio return, annualized by $\sqrt{52}$. The high-minus-low portfolio uses all-stock breakpoints, the baseline AI beta, and the same value-weighted specification as the baseline portfolio-sort tests. Data is weekly from January 2024 through April 2026.

Table IA.13: Spanning Regressions for the AI-Beta H−L Portfolio with High-Tech Returns

	α	$t(\alpha)$	b_{Add}	$t(b_{\text{Add}})$	$\bar{R}^2$
Panel A. FF10 HiTec return as additional control					
Value-weighted, all-stock breakpoints					
(1) HiTec	0.625	(2.74)	0.043	(0.60)	0.004
(2) Mkt + HiTec	0.603	(2.69)	-0.351	(-1.73)	0.048
(3) FF5 + HiTec	0.566	(2.43)	-0.160	(-0.43)	0.063
(4) FF5 + Mom + HiTec	0.562	(2.42)	-0.201	(-0.54)	0.075
Value-weighted, NYSE breakpoints					
(1) HiTec	0.524	(2.44)	0.025	(0.37)	0.001
(2) Mkt + HiTec	0.498	(2.39)	-0.419	(-2.22)	0.065
(3) FF5 + HiTec	0.477	(2.20)	-0.348	(-1.01)	0.075
(4) FF5 + Mom + HiTec	0.472	(2.19)	-0.396	(-1.14)	0.093
Panel B. NAICS semiconductor return as additional control					
Value-weighted, all-stock breakpoints					
(1) Semi	0.635	(2.75)	0.005	(0.13)	0.000
(2) Mkt + Semi	0.611	(2.69)	-0.112	(-1.59)	0.044
(3) FF5 + Semi	0.568	(2.41)	-0.044	(-0.49)	0.064
(4) FF5 + Mom + Semi	0.573	(2.44)	-0.078	(-0.84)	0.079
Value-weighted, NYSE breakpoints					
(1) Semi	0.534	(2.46)	-0.001	(-0.03)	0.000
(2) Mkt + Semi	0.510	(2.39)	-0.116	(-1.75)	0.047
(3) FF5 + Semi	0.481	(2.18)	-0.051	(-0.61)	0.069
(4) FF5 + Mom + Semi	0.485	(2.21)	-0.090	(-1.03)	0.091
Panel C. AI/robotics ETF basket as additional control					
Value-weighted, all-stock breakpoints					
(1) AI ETF	0.630	(2.77)	0.032	(0.49)	0.003
(2) Mkt + AI ETF	0.588	(2.62)	-0.294	(-1.80)	0.051
(3) FF5 + AI ETF	0.555	(2.40)	-0.232	(-1.03)	0.072
(4) FF5 + Mom + AI ETF	0.551	(2.38)	-0.234	(-1.04)	0.083
Value-weighted, NYSE breakpoints					
(1) AI ETF	0.521	(2.44)	0.035	(0.58)	0.004
(2) Mkt + AI ETF	0.489	(2.30)	-0.211	(-1.36)	0.035
(3) FF5 + AI ETF	0.465	(2.14)	-0.175	(-0.83)	0.072
(4) FF5 + Mom + AI ETF	0.461	(2.13)	-0.176	(-0.84)	0.087

The dependent variable is the H−L value-weighted AI-beta quintile spread. Panel A adds the Fama-French 10-industry HiTec return, Panel B adds the value-weighted weekly return on the NAICS-defined semiconductor portfolio, and Panel C adds the simple weekly excess return on an equal-weighted basket of seven AI ETFs (BOTZ, AIQ, IRBO, ROBO, ARTY, WTAI, CHAT). The ETF basket return is the simple-return aggregation of the daily basket minus the Fama-French weekly risk-free rate. b_{Add} is the coefficient on the panel-specific additional control. *Mkt* is the Fama-French weekly market excess return. FF5 is the Fama-French five-factor model (Fama and French, 2015), and FF5+Mom adds the momentum factor (Carhart, 1997). Each row is one OLS spanning regression, and no row combines multiple additional controls. Data is weekly from January 2024 through April 2026.

Table IA.14: China A-Share Quintile Spreads Sorted on AI-Consumption Betas

	Raw		China-market α	
Weight	α	t	α	t
Panel A. China A-shares sorted on β^{AI}				
Value-weighted	-0.159	(-0.61)	-0.167	(-0.63)
Equal-weighted	-0.072	(-0.37)	-0.073	(-0.36)
Panel B. China A-shares sorted on $\beta^{AI,China}$				
Value-weighted	-0.241	(-1.03)	-0.113	(-0.45)
Equal-weighted	-0.300	(-1.59)	-0.219	(-1.07)

The table reports value-weighted and equal-weighted High−Low spreads for China A-share stocks sorted into pooled quintiles by their 13-week rolling log-excess AI beta. Panel A uses the overall AI factor. Panel B uses a China-specific AI factor constructed from the subset of AI consumption originating from users in mainland China. The sample is the Datastream China A-share common-stock panel. *Raw* is the time-series mean of the H−L spread. *China-market α* is the intercept from regressing the H−L spread on the value-weighted Datastream A-share market log-excess return. Returns and alphas are in percent per week. Data is weekly from January 2024 through April 2026.

IA.4 Event Study of AI Model Releases

This appendix reports the details behind the event study in Section 3.3.4, where high-β^{AI} firms are shown to outperform low-β^{AI} firms in narrow windows around major AI model releases. We classify releases into two groups by provider. Frontier providers are Anthropic, DeepSeek, Google, Meta, and OpenAI, and yield 19 events in the sample; releases from other providers add 28 more. Table IA.16 documents each event with its sources.

Table IA.15 reports the CAR comparison in statistical form. Panel A uses frontier-provider releases and Panel B uses releases from other providers, each over three event windows ($(-1,+1)$, $(0,+1)$, and $(0,+5)$) and adjusted for the market model, the Fama-French three- and five-factor models, and FF5 plus momentum. The frontier-provider effect dominates throughout. In Panel A, the value-weighted CAR difference over the $(0,+5)$ window is 1.933 percent (1.071 percent after FF5+Mom adjustment), against 0.905 percent for non-frontier releases (and weaker after the same controls).

Table IA.15: AI-Beta Long-Short Portfolio Returns Around AI Model Releases

Panel A. Frontier model releases (19 events)			
	$(-1,+1)$	$(0,+1)$	$(0,+5)$
Raw	1.284	0.876	1.933
t	(3.36)	(2.63)	(4.17)
Market model	1.364	0.930	1.355
t	(3.25)	(2.73)	(2.49)
FF3	1.085	0.710	0.836
t	(2.34)	(2.15)	(1.57)
FF5	1.114	0.836	1.046
t	(2.41)	(2.63)	(2.03)
FF5+Mom	0.703	0.587	1.071
t	(1.98)	(2.14)	(2.20)
Panel B. Non-frontier model releases (28 events)			
	$(-1,+1)$	$(0,+1)$	$(0,+5)$
Raw	0.482	0.452	0.905
t	(1.37)	(1.75)	(2.24)
Market model	0.373	0.284	0.724
t	(0.98)	(1.05)	(1.62)
FF3	0.361	0.290	0.647
t	(1.36)	(1.75)	(1.91)
FF5	0.195	0.271	0.568
t	(0.69)	(1.58)	(1.67)
FF5+Mom	0.340	0.288	0.480
t	(1.38)	(2.04)	(1.41)

Panel A uses 19 frontier AI model release dates from Anthropic, OpenAI, Google, Meta, and DeepSeek. Panel B uses 28 verified non-frontier releases from non-top-5 providers. For each event, firms in the eligible U.S. universe are sorted by pre-event 13-week rolling AI beta and formed into value-weighted High-minus-Low portfolios. Windows are trading days from the event anchor. *Raw* is the cumulative simple return. The Market model, Fama–French three-factor model (FF3; Fama and French, 1993), Fama–French five-factor model (FF5; Fama and French, 2015), and five-factor model augmented with momentum (FF5+Mom; Carhart, 1997) rows subtract expected returns from the corresponding model estimated over the prior $(-250,-30)$ trading-day window. Reported numbers are cross-event mean H−L returns in percent, with *t*-statistics in parentheses. Data is weekly from January 2024 through April 2026.

Table IA.16: AI Model Release Calendar and Primary Sources

Panel A. Frontier model release calendar (19 events)

Date	Provider	Model	Source	Note
2024-04-18	Meta	Llama 3	ai.meta.com	Open model release; strong open-weight benchmark step.
2024-05-13	OpenAI	GPT-4o	openai.com	OpenAI flagship omni model; cheaper and faster API.
2024-05-14	Google	Gemini 1.5 Flash	blog.google	Fast/efficient Gemini family update at Google I/O.
2024-06-21	Anthropic	Claude 3.5 Sonnet	anthropic.com	Anthropic frontier Sonnet model.
2024-07-23	Meta	Llama 3.1 405B	ai.meta.com	First frontier-level open-source AI model per Meta.
2024-09-12	OpenAI	o1-preview	openai.com	OpenAI first public reasoning-model series.
2024-10-22	Anthropic	Claude 3.5 Sonnet new	anthropic.com	Upgraded Sonnet plus public beta computer use.
2024-12-06	Meta	Llama 3.3 70B	ai.meta.com	Efficiency/cost-improved Llama model; source is Meta year-end review.
2024-12-11	Google	Gemini 2.0	blog.google	Google agentic-era model release.
2024-12-27	DeepSeek	DeepSeek-V3	arxiv.org	Open-weight frontier MoE model; arXiv date used for public paper/checkpoint availability.
2025-01-20	DeepSeek	DeepSeek-R1 release	docs.aws.amazon.com	Official model launch date per AWS Bedrock model card.
2025-01-27	DeepSeek	DeepSeek-R1 market recognition	axios.com	US market-recognition shock one week after the Jan. 20 R1 release.
2025-02-24	Anthropic	Claude 3.7 Sonnet	anthropic.com	Hybrid reasoning model.
2025-02-27	OpenAI	GPT-4.5	openai.com	OpenAI largest GPT model research preview.
2025-03-25	Google	Gemini 2.5 Pro	blog.google	Google thinking model.
2025-04-05	Meta	Llama 4 Scout/Maverick	ai.meta.com	Open-weight natively multimodal Llama 4 models.
2025-04-14	OpenAI	GPT-4.1	openai.com	API model with coding, instruction-following, and long-context gains.
2025-04-16	OpenAI	o3 and o4-mini	openai.com	OpenAI o-series reasoning models with tool access.
2025-05-22	Anthropic	Claude 4	anthropic.com	Claude Opus 4 and Sonnet 4.
2025-08-07	OpenAI	GPT-5	openai.com	OpenAI unified GPT-5 system.

Panel B. Non-frontier model release calendar (28 events)

Date	Provider	Model	Source	Note
2024-03-11	Cohere	Command R	cohere.com	RAG / tool-use enterprise LLM (128K context).
2024-03-17	xAI	Grok-1 (open weights)	x.ai	314B-parameter MoE base model, Apache 2.0.
2024-03-27	Databricks	DBRX	databricks.com	132B-parameter MoE open LLM (36B active).
2024-03-28	AI21	Jamba	ai21.com	First production-grade hybrid SSM-Transformer LLM.
2024-04-04	Cohere	Command R+	cohere.com	Scaled-up RAG/tool-use enterprise LLM (104B).
2024-04-15	Reka	Reka Core	reka.ai	Frontier-class multimodal model (text/image/video/audio).
2024-04-23	Microsoft	Phi-3 (mini)	azure.microsoft.com	3.8B small language model launched on Azure.
2024-05-13	01.AI	Yi-Large	x.com	01.AI's first proprietary closed-source frontier dense LLM.
2024-06-05	Zhipu	GLM-4-9B (open weights)	github.com	Open-weight 9B variant of the GLM-4 family.
2024-06-07	Alibaba	Qwen 2	qwenlm.github.io	Qwen 2 family (0.5B-72B) open-source release.
2024-07-24	Mistral	Mistral Large 2	mistral.ai	123B-parameter flagship multilingual / code-strong dense LLM.
2024-08-01	Black Forest Labs	FLUX.1	bfl.ai	12B rectified-flow text-to-image model family.
2024-08-13	xAI	Grok-2	x.ai	Grok-2 / Grok-2 mini early-preview release on X.
2024-08-22	AI21	Jamba 1.5 family	ai21.com	Hybrid SSM-Transformer family with 256K context.
2024-09-19	Alibaba	Qwen 2.5	qwenlm.github.io	Open-source family announced at Apsara Conference.
2024-10-21	IBM	Granite 3.0	newsroom.ibm.com	Apache-2.0 enterprise LLM family launched at TechXchange.
2024-11-05	Tencent	Hunyuan-Large	huggingface.co	389B-parameter MoE (52B active) open-sourced.
2024-11-18	Mistral	Pixtral Large	mistral.ai	124B open-weights multimodal model (123B + 1B vision).
2024-11-28	Alibaba	QwQ-32B-Preview	qwenlm.github.io	Open-weight reasoning model preview from Qwen.
2024-12-03	Amazon	Nova family	press.aboutamazon.com	Foundation-model family launched on Amazon Bedrock at re:Invent.
2024-12-12	Microsoft	Phi-4	huggingface.co	14B reasoning-focused SLM (Azure preview, fully open Jan 2025).
2025-01-22	ByteDance	Doubao 1.5 Pro	seed.bytedance.com	Large-scale sparse MoE flagship on Volcano Engine.
2025-01-28	Alibaba	Qwen 2.5-Max	qwenlm.github.io	Closed-source MoE flagship API model.
2025-02-17	xAI	Grok-3	x.ai	Flagship reasoning model trained on Colossus.
2025-03-13	Cohere	Command A	cohere.com	111B enterprise model (256K context, 2 GPU inference).
2025-04-29	Alibaba	Qwen 3	qwenlm.github.io	Hybrid thinking/non-thinking dense + MoE family.
2025-06-10	Mistral	Magistral (Small + Medium)	mistral.ai	Mistral's first reasoning-focused models.
2025-07-09	xAI	Grok 4	x.ai	Grok 4 / Grok 4 Heavy livestream release.

Dates are announced public release dates and sources are primary citations. Panel A is the frontier-event sample. Panel B excludes Anthropic, OpenAI, Google, Meta, and DeepSeek and keeps flagship or series-defining releases. Sample period: January 2024 through April 2026.

IA.5 AI Exposure

This section reports the industry-level exposure map behind the firm-level and portfolio evidence. Figure IA.5 shows how the baseline AI factor and the split-specific AI factors revalue the Fama-French 10 industries. Each cell reports the value-weighted average firm-level AI beta within an industry, scaled by the weekly standard deviation of the relevant AI factor, so the numbers can be read as the percentage-point response of an industry-average weekly return to a one-standard-deviation shock to that AI factor. Most industries have positive baseline AI exposure, with the largest values in retail and consumer durables. Nondurable goods and health are the only industries with negative exposure on average. The high-technology sector is only modestly exposed on the baseline factor, masking strongly positive exposure of semiconductor and AI-infrastructure firms and negative exposure elsewhere in technology. The cross-factor columns show that exposures are larger and more concentrated when the factor is built from closed-source models, paid/core users, seasoned users, and long prompts. The figure reads AI exposure as a broad diffusion and deployment shock rather than as a narrow technology-sector label.

Figure IA.5: Industry Exposure to AI Factor



The figure reports the implied weekly industry return in percent associated with a one-standard-deviation increase in each AI variable for the Fama-French 10 industries. For each firm-week, the relevant beta is estimated from the same 13-week rolling regression of weekly log-excess returns on the log-excess market return and the AI variable used in the portfolio sorts. Within each FF10 industry, we take the value-weighted mean of firm-level betas using formation-week market equity weights; each cell reports $100 \times \bar{\beta}^{X}_{j,\mathrm{vw}} \times \sigma_w(X)$ for the column variable X. The baseline AI factor is the first principal component of weekly log growth in total tokens, usage dollars, and distinct users. Panel A compares AI, AI^{Closed}, and AI^{Open}. Panel B compares AI, AI^{Core}, and $AI^{Noncore}$. Panel C compares AI, $AI^{Seasoned}$, and $AI^{Nonseasoned}$. Panel D compares AI, $\Delta \ln Tok^{Long}$, and $\Delta \ln Tok^{Short}$. Samples vary by split according to non-missing rolling-beta availability. Cell shading is scaled within each column; blue denotes positive responses and red denotes negative responses. Data is weekly from January 2024 through April 2026.

IA.6 Additional AI Stylized Facts

This section collects additional stylized facts on AI consumption based on the OpenRouter data.

Table IA.17 lists the agentic-economy concepts we extract from the OpenRouter data and the weekly factor that proxies each. These salient components of AI token consumption associated with agentic calls are used in the analysis of firms' exposure to agentic consumption discussed in Section 5 of the paper and specifically in Table 11.

Table IA.17: Agentic-Economy Concepts and Factor Definitions

Factor	Concept	Construction
Agentic tokens	Tool use	Log growth in total tokens on requests whose normalized finish reason is `tool_calls`.
Cache-read tokens	Context reuse	Log growth in cache-read tokens—prompt-prefix tokens served from cache rather than recomputed.
Reasoning tokens	Test-time compute	Log growth in reasoning tokens—the model's internal reasoning ("work plan") tokens.
Agentic dollars	Token economics	Log growth in usage dollars on requests whose normalized finish reason is `tool_calls`.
Composite (PC1)	Agentic intensity	First principal component of the four standardized growth factors above.

Each row is one weekly factor built from the OpenRouter inference logs; the corresponding token and dollar shares are plotted in Figure 7. Total tokens are prompt plus completion tokens. Each of the first four factors is a weekly log growth, $\ln L_t - \ln L_{t-1}$, of a positive weekly level; the composite is the first principal component of the four standardized growth series. The agentic, cache-read, and reasoning slices are populated only after their instrumentation onset.

Table IA.18 compares the granular OpenRouter user–model–day panel with the scraped, aggregate data from the OpenRouter public website used by Demirer, Fradkin, Tadelis and Peng (2025). Their public scrape records posted prices and aggregate token counts but cannot observe realized spending, individual users, or within-model variation, and four of their headline findings change once these margins are visible in our licensed user–model–day panel. First, Demirer, Fradkin, Tadelis and Peng (2025) report that the average price per token is roughly constant. The series they recover is a token-weighted list price, which the scrape cannot distinguish from the price actually paid. We find that the price paid fell by about 30 percent within their 2025 window and by about half over our full sample. Second, Demirer, Fradkin, Tadelis and Peng (2025) put the open-weight share at about 48 percent of tokens. The same models earn only about 15 percent of realized dollars, overstating open-weight's economic weight by roughly threefold. Third, Demirer, Fradkin, Tadelis and Peng (2025) document rapid turnover among leading models and creators. Weighted by realized dollars rather than tokens, the market instead consolidates: creator-level concentration rises by 72 percent. Fourth, Demirer, Fradkin, Tadelis and Peng (2025) estimate short-run prompt-price elasticities just above one. With model-date and provider fixed effects in our panel, these elasticities become small and statistically insignificant. The table also reports application-level multihoming: only 35.8 percent of applications use at least two models in the latest month, with the median top-model token share at 100 percent.

Table IA.18: Granular OpenRouter Facts for Comparison with Demirer, Fradkin, Tadelis and Peng (2025)

Panel A. Open-Source Token and Dollar Shares

Period	Week	Class	Request share	Token share	Dollar share
First week	2024-01-01	Closed	9.4	15.6	51.9
First week	2024-01-01	Open	90.6	84.4	48.1
Final week	2026-04-13	Closed	54.8	51.6	84.7
Final week	2026-04-13	Open	45.2	48.4	15.3

Panel B. Market Dynamism in Tokens and Dollars

Metric	Model HHI first	Model HHI latest	Change	Author HHI first	Author HHI latest	Change
Requests	1321.7	267.5	-79.8	1764.7	1565.1	-11.3
Tokens	1057.9	333.4	-68.5	1508.6	1117.0	-26.0
Usage dollars	975.6	959.3	-1.7	1996.2	3435.8	72.1

Metric	Horizon	Model top-10 retention	Author top-10 retention
Requests	4 weeks	87.0	91.0
Requests	8 weeks	85.0	90.0
Requests	12 weeks	80.0	90.0
Tokens	4 weeks	58.0	78.0
Tokens	8 weeks	36.7	73.3
Tokens	12 weeks	35.0	85.0
Usage dollars	4 weeks	70.0	92.0
Usage dollars	8 weeks	50.0	86.7
Usage dollars	12 weeks	35.0	80.0

Panel C. Prompt-Price Elasticity Comparison

Specification	Paper coef.	Our coef.	Our t-stat.	Obs.
Date FE	-0.55	-0.450	(-4.67)	52,728
Date + model FE	-1.08	-1.170	(-5.66)	52,728
Model-date + model-provider FE	-1.11	0.280	(0.79)	53,800

Panel D. Multihoming and Concentration Across OpenRouter Entities

Entity	First	Latest	≥2 first	≥2 latest
API key	2025-01-01	2026-04-01	48.7	51.8
Application	2025-01-01	2026-04-01	38.3	35.8

Panel A reports shares among the open-plus-closed model universe using realized requests, prompt-plus-completion tokens, and usage dollars. Panel B reports Herfindahl-Hirschman indices across models and authors and average top-10 retention in the latest 90 days of the sample. Panel C compares the prompt-price elasticity estimates in Demirer, Fradkin, Tadelis and Peng (2025), Table 2, with the corresponding granular OpenRouter reconstruction. The sample is open-source model-provider-days, the dependent variable is log token quantity, and the price regressor is log posted prompt price. Panel D reports multihoming among OpenRouter applications and API keys. Usage dollars are the realized OpenRouter `usage` field, not a recomputation from posted prices. Shares and changes are reported in percent. Data is weekly from January 2024 through April 2026.

IA.7 Skill Exposure to AI

This section reports the occupation-level exposure map that underlies the skill-exposure exercise in the main text. Figure IA.6 maps firm and industry AI exposures into SOC minor occupation groups using BLS occupation-by-industry employment weights, and plots the resulting employment-weighted occupation exposure to the baseline AI factor. Panels A and B together rank all 109 SOC minor occupation groups, with the colored rail flagging coherent economic themes. Positive exposure clusters in installation, maintenance, programming, persuasion, instruction, and systems-integration occupations. Negative exposure clusters in science, healthcare, and operations-control occupations. The pattern is the occupation-level counterpart of the firm-level finding that the AI premium is broad rather than confined to the technology sector.

Panel A. Ranks 1–55



Panel B. Ranks 56–109



Figure IA.6: Occupation Exposure to the AI Factor

The figure plots the full-firm occupation ranking based on SOC minor occupation groups. NAICS-level exposure is the market-equity-weighted mean of firm-level AI betas within NAICS-week cells, requiring at least three firms in a NAICS-week, averaged over time and mapped to occupations using BLS OEWS occupation-by-industry employment weights. Detailed SOC occupations are aggregated to SOC minor groups using employment weights. The plotted value is $100 \times \beta_o^{AI}$, where β_o^{AI} is the employment-weighted occupation exposure to the baseline AI factor. The colored rail marks the coherent economic theme assigned to each occupation group. Panel A reports ranks 1–55, and Panel B reports ranks 56–109. Underlying firm-level AI betas are weekly from January 2024 through April 2026.

IA.8 Variable Definitions

Table IA.19 reports the definitions of the variables used in the analysis and their correspondence in the OpenRouter dataset and the additional datasets and sources used in the paper.

Table IA.19: Variable Definitions

Name	Definition	Dataset used
AI consumption and attention variables		
Tok_t	Total OpenRouter prompt plus completion tokens in week t.	OpenRouter
Dol_t	Total dollar-denominated OpenRouter usage in week t.	OpenRouter
$User_t$	Distinct active OpenRouter users in week t.	OpenRouter
$\Delta \ln Tok_t$	Weekly log growth in total OpenRouter tokens.	OpenRouter
$\Delta \ln Dol_t$	Weekly log growth in OpenRouter usage dollars.	OpenRouter
$\Delta \ln User_t$	Weekly log growth in distinct active OpenRouter users.	OpenRouter
AI_t	First principal component of $\Delta \ln Tok_t$, $\Delta \ln Dol_t$, and $\Delta \ln User_t$.	OpenRouter
AI_t^{Closed}, AI_t^{Open}	AI factors constructed separately from closed-weight and open-weight model consumption.	OpenRouter; model catalog
AI_t^{Core}, AI_t^{New}	AI factors constructed separately from paid/core-user and new-user consumption.	OpenRouter account data
$AI_t^{Noncore}$	AI factor constructed from non-paid/core account consumption, the complement of paid/core account consumption.	OpenRouter account data
$AI_t^{Seasoned}$, $AI_t^{Nonseasoned}$	AI factors constructed separately from seasoned-user and non-seasoned-user consumption.	OpenRouter account data
$\Delta \ln Tok_t^{Long}$, $\Delta \ln Tok_t^{Short}$	Weekly token-growth series for long prompts and short prompts, split by the rolling prompt-length cutoff.	OpenRouter row-level usage
$Share_{c,t}^{cat}$	Share of categorized tokens assigned to prompt category c in week t.	OpenRouter category-token file
$Share_t^{Closed}$, $Share_t^{Core}$, $Share_t^{Seasoned}$, $Share_t^{Long}$	Weekly usage shares for closed-weight models, paid/core users, seasoned users, and long prompts.	OpenRouter; model catalog; account data
$\Delta \ln \text{gtrend}_t^{AI}$	Weekly log growth in Google Trends interest for the topic "Artificial intelligence."	Google Trends
Return, exposure, and control variables		
$r_{i,t}^e$	Weekly log excess return of stock i over the risk-free rate.	CRSP; Fama-French
$r_{m,t}^e$ or rm_t	Weekly log excess market return.	Fama-French
$\beta_{i,t}^{rm}$	Rolling market beta of stock i from the first-pass return regression.	CRSP; Fama-French
$\beta_{i,t}^{AI}$	Rolling AI exposure of stock i: coefficient on the AI factor in a 13-week regression of $r_{i,t}^e$ on $r_{m,t}^e$ and AI_t.	CRSP; OpenRouter; Fama-French

Name	Definition	Dataset used
$\beta_{i,t}^{Tok}$	Rolling exposure of stock i to $\Delta \ln Tok_t$.	CRSP; OpenRouter; Fama-French
$\beta_{i,t}^{Dol}$	Rolling exposure of stock i to $\Delta \ln Dol_t$.	CRSP; OpenRouter; Fama-French
$\beta_{i,t}^{User}$	Rolling exposure of stock i to $\Delta \ln User_t$.	CRSP; OpenRouter; Fama-French
$\beta_{i,t}^{AI,Ind}$	AI beta minus the same-week median AI beta of stock i's Fama-French 30 industry.	CRSP; Compustat; OpenRouter; Fama-French
$r^{e}_{\text{HiTec},t}$	Fama-French HiTec industry excess return used as a first-pass control.	Fama-French
$r_{\text{semi},t}$	Weekly log return of the U.S. semiconductor industry used as a first-pass control.	CRSP; Compustat
$\beta_{i,t}^{AI,Closed}$, $\beta_{i,t}^{AI,Open}$	Rolling exposures of stock i to AI_t^{Closed} and AI_t^{Open}.	CRSP; OpenRouter; model catalog
$\beta_{i,t}^{AI,Core}$, $\beta_{i,t}^{AI,New}$	Rolling exposures of stock i to AI_t^{Core} and AI_t^{New}.	CRSP; OpenRouter account data
$\beta_{i,t}^{AI,Seasoned}$, $\beta_{i,t}^{AI,Nonseasoned}$	Rolling exposures of stock i to $AI_t^{Seasoned}$ and $AI_t^{Nonseasoned}$.	CRSP; OpenRouter account data
$\beta_{i,t}^{Tok,Long}$, $\beta_{i,t}^{Tok,Short}$	Rolling exposures of stock i to $\Delta \ln Tok_t^{Long}$ and $\Delta \ln Tok_t^{Short}$.	CRSP; OpenRouter row-level usage
$\ln ME_{i,t}$	Natural log of stock i's formation-week market equity.	CRSP; Compustat
$\ln BM_{i,t}$	Natural log of stock i's book-to-market equity.	Compustat; CRSP
$Profit_{i,t}$	Gross profitability, revenue minus cost of goods sold scaled by assets.	Compustat
$Inv_{i,t}$	Asset growth, measured as the year-over-year change in total assets scaled by lagged assets.	Compustat
$Mom_{i,t}$	Prior stock return from week $t-52$ through week $t-5$.	CRSP
$Rev_{i,t}$	Prior one-month stock return used as a short-term reversal control.	CRSP
$Lev_{i,t}$	Long-term debt scaled by assets.	Compustat
$Accr_{i,t}$	Accruals.	Compustat
International and event-study variables		
$r_{i,t}^{local}$	Weekly local-currency return of non-U.S. stock i.	Compustat Global
$r_{m,t}^{local}$	Weekly regional local-market return.	Compustat Global; market returns
AI_t^{China}	AI factor constructed from OpenRouter usage originating from mainland China.	OpenRouter country-coded usage
$r_{m,t}^{China}$	Weekly Datastream China A-share market log excess return.	Datastream; Fama-French
r_t^{AIETF}	Weekly excess return on the AI and robotics ETF basket used as a first-pass and spanning control.	ETF return data; Fama-French
$CAR_{e,[a,b]}$	Cumulative abnormal H-minus-Low return for event e over trading-day window $[a,b]$.	CRSP; Fama-French; release calendar
$CAAR_{[a,b]}$	Cross-event average of $CAR_{e,[a,b]}$ over the specified event window.	Constructed