Capable but Not Deployable: Institutional Constraints on AI Exposure in Finance

Claes Bäckman, Christos A. Makridis, and Lisa Simon *

November 2, 2026

Abstract

Empirical measures of AI exposure ask language models to score O*NET tasks for technical feasibility. In finance, technically feasible tasks must still pass through review, documentation, supervision, confidentiality controls, and accountable human sign-off before entering production. We measure the gap between feasibility and institutional deployability using 2,199 O*NET tasks across 99 finance-and-insurance occupations. We score each task with eight frontier models and a prompt ladder that moves from bare capability to finance-industry context and named regulatory regimes. The within-model institutional markdown is about one-fifth of the mean feasibility score, and positive for all eight models. The markdown is largest for regulated, client-facing credit and advice roles and smallest for marketing, software, and support roles. Cross-model agreement also declines as finance context is added: models agree more about what AI can do than about what financial institutions can deploy. Mapping exposure to publicly traded firms through pre-ChatGPT staffing shares, we find that the pricing content resides in the institutional layer: firms in the top half of the markdown distribution underperform the bottom half by roughly 25 percentage points in market-adjusted cumulative abnormal returns over the three years after ChatGPT, while sorting on technical exposure alone produces no gap. The differential lies outside the range the same design produces over every pre-ChatGPT window of equal length, though with one event window and few subsector clusters we read it as evidence on where return information resides rather than as a causal estimate. Especially in regulated industries, deployable exposure rather than technical feasibility is the more relevant measure of AI exposure.

Keywords: AI exposure; large language models; occupational tasks; finance labor markets; technology adoption; firm exposure; asset pricing.
JEL Codes: C52; C81; G12; G21; J23; J24; O33.

*Bäckman: Leibniz Institute for Financial Research SAFE, claes.backman@gmail.com. Makridis: Arizona State University, Stanford University, Gallup Inc., cmakridi@asu.edu; Simon: Revelio Labs, lisa@reveliolabs.com. The authors used Claude and Claude Code heavily (and some ChatGPT) to automate the data collection, analysis, and writing for the purposes of the AFA call for papers around AI-guided papers. After using these tools, the authors reviewed and edited the content as needed and take responsibility. These views are our own and do not reflect those of any affiliated institutions.

1 Introduction

Large language models can now perform a wide range of tasks used in modern knowledge work. They can read and summarize complex documents, draft communications, extract and classify information, write code, and produce first-pass analysis.[1] Technical capability and deployment are separate objects. A model may complete a task when it is evaluated in isolation, yet the output may never enter a firm's production process because adoption also depends on task-specific user costs, worker comparative advantage, complementary software and data, managerial strategy, and the rules governing how outputs are reviewed and used. Task-based exposure measures therefore map the technological possibility set; they do not, by themselves, identify which capabilities firms will actually put into production. Realized deployment instead depends on comparative advantage and user costs (Lindenlaub et al., 2026), organizational complementarities (Makridis, 2026), and sector-specific regulatory constraints (Makridis et al., 2026; Parshall and Lopez-Luzuriaga, 2026).

Finance provides a useful setting in which to study the resulting wedge. LLM exposure is concentrated in cognitive, language-intensive, and relatively high-wage work (Eloundou et al., 2024; Cazzaniga et al., 2024), and financial services employ a dense mix of professional occupations organized around exactly these activities. Financial workers read complex records, synthesize information, classify cases, forecast outcomes, explain decisions, and communicate with clients. Existing measures consequently place financial, legal, mathematical, and administrative occupations near the frontier of generative-AI exposure (Eisfeldt et al., 2023), while evidence on actual

[1] A growing experimental and field literature documents the effects of generative AI across professional writing (Noy and Zhang, 2023), management consulting (Dell'Acqua et al., 2026), customer support (Brynjolfsson et al., 2025), and software development (Peng et al., 2023); see also Eloundou et al. (2024) and Felten et al. (2021) for task- and ability-based assessments of occupational exposure.

use shows rapid adoption within finance and related professional work (Makridis et al., 2026).

While finance should technically be among the sectors in which LLMs have the greatest scope to affect production, technically feasible outputs in finance also face unusually demanding conditions before they can be deployed. A model can summarize a loan file, draft a client email, classify a suspicious-transaction alert, or prepare a first-pass investment note. The economic question is not whether the model can produce a plausible artifact, but rather whether that artifact can enter production once the institution must preserve records, protect customer data, explain decisions, supervise communications, validate recurring models, and keep an accountable human in the loop. This combination of high technical exposure and strong institutional constraints makes finance a natural setting for separating what LLMs can do from what firms can use.

This paper measures that distinction and quantifies the implications. The standard empirical approach begins with task content, scoring O*NET tasks for the extent to which AI can reduce the time required to complete them (Eloundou et al., 2024), building on earlier occupation- and technology-linkage measures (Felten et al., 2021; Webb, 2020). We retain that task-based structure but score two different objects: bare technical feasibility and deployability inside a financial institution. Holding the task, model, and score scale fixed, we define the difference between them as the institutional markdown. Across eight frontier-model checkpoints, the markdown is large, appears throughout the model panel, and is concentrated in regulated and client-facing occupations. The distinction also matters for downstream finance applications. Feasibility-based scores are increasingly aggregated to the firm level and used to study labor demand and asset prices (Eisfeldt et al., 2023); if cash flows respond to capabilities firms can deploy rather than capabilities models possess in isolation, then unadjusted exposure measures misstate the relevant object.

We implement the idea in finance because its deployment constraints are clearly defined, salient,

and large (Makridis et al., 2026). Federal Reserve model-risk guidance, ECOA and Regulation B adverse-action rules, GLBA confidentiality obligations, FINRA and SEC supervisory-review and recordkeeping requirements, and accountable-human duties are not generic organizational frictions,[2] specifying when an AI output can be used, reviewed, explained, or acted on. The same technical capability therefore has different deployable content when the task is performed inside a bank, broker-dealer, insurer, investment adviser, payments firm, or other NAICS 52 institution.

Our measurement uses a prompt ladder. For each task–model pair, the model estimates, on a 0–100 scale, the share of human task time that a frontier AI system could reduce or eliminate while preserving the quality standard the task requires. We score 2,199 O*NET task statements across 99 finance-and-insurance occupations using eight model checkpoints from the GPT and Claude families. Rung A asks the model to assess bare technical feasibility, abstracting from organizational and institutional constraints. Rung B then places the same task inside a financial institution, allowing the model to account for general requirements such as review, accountability, confidentiality, and workflow integration. Rung C makes the institutional environment more explicit by naming the relevant finance-governance domains, including explainability, recordkeeping, model validation, data protection, and accountable human oversight. We define the $A - C$ difference as the institutional markdown, or deployment haircut, and decompose it into an $A - B$ finance-context adjustment and a $B - C$ named-governance adjustment. Because the task, model

[2]The specific anchors are as follows. Federal Reserve SR 11-7 and OCC Bulletin 2011-12 require banks to inventory, validate, and monitor models used in material decisions. The Equal Credit Opportunity Act and Regulation B (12 C.F.R. §1002) require creditors to give applicants specific, accurate reasons for adverse credit actions. The Gramm-Leach-Bliley Act privacy and safeguards rules restrict how nonpublic customer information can be shared and processed. FINRA Rules 2210 and 3110 and SEC Rules 17a-3 and 17a-4 require supervisory review, approval, and multi-year retention of communications and records at broker-dealers, with parallel marketing and books-and-records rules for investment advisers. Regulation Best Interest, fiduciary duties, and SOX §§302 and 404 certification requirements keep a named human accountable for recommendations and financial reports. Appendix A maps each regime to a coded constraint dimension.

checkpoint, score scale, and output schema remain fixed across rungs, these differences capture how the same model revises its assessment when institutional context is introduced, rather than differences in how models use the numerical scale. Applying the same ladder across eight checkpoints also allows us to assess whether the institutional markdown is driven by a particular model or appears across both model families—and it is common across model families.

The main result is that the institutional markdown is large. The employment-weighted markdown is a statistically significant 12.0 points on a 0–100 exposure scale against a mean rung-A feasibility score of 62 points. Roughly one-fifth of measured technical exposure disappears once the task is evaluated inside finance. Every model has a positive markdown, with point estimates ranging from 1.3 to 18.6 points in the importance-weighted estimates. Most of the adjustment occurs when the task is first placed inside finance rather than when named regimes are added: the average $A - B$ adjustment is 10.1 points, compared with 1.9 points for $B - C$. Models reprice the task as soon as they learn where it is performed, before any specific regulation is named.

The markdown is concentrated where the construct predicts it should be. On one hand, loan interviewers, credit analysts, credit authorizers, new-account clerks, loan officers, and insurance underwriters sit at the top of the distribution. These are high-feasibility occupations whose work is tied to customer decisions, sensitive data, regulated communications, or accountable review. On the other hand, market-research analysts, graphic designers, software developers, web-interface designers, and maintenance workers sit near the bottom. Some of these occupations are technically exposed, but their tasks are less often tied to credit decisions, customer files, or supervised financial advice. Within finance, our results provide evidence for a "capable but not deployable" region: occupations with high rung-A technical feasibility also tend to lose more exposure once the task is embedded in finance. A firm staffed with high-feasibility, high-markdown occupations is not

exposed to AI in the same way as a firm staffed with high-feasibility, low-markdown occupations, and the two firms should not receive the same exposure score.

The measure also behaves sensibly against external benchmarks. Our rung-A feasibility score tracks the Eloundou et al. (2024) task-exposure measure when both are aggregated over the same tasks and O*NET weights: the occupation-level correlation is 0.73 against their GPT-4 rating and 0.53 against their human annotation. The markdown is a different object. It correlates only 0.23 with the same benchmark at the occupation level, so the institutional layer captures information that feasibility-based measures abstract from. If anything, the measured markdown understates the wedge between context-free feasibility and deployable exposure: although the rung-A prompt never names an industry, on average 17 percent of the models' free-text rationales per task already invoke finance-institutional language, so part of the institutional signal is embedded in the feasibility rung itself.[3]

Institutional context lowers measured exposure and reduces cross-model agreement: cross-model rank agreement is highest at rung A (0.804 mean pairwise Spearman correlation) and declines as institutional context is added (0.738 at rung C). The omitted institutional layer is both a level adjustment and a source of instability in the exposure construct.

Separating feasibility from deployability also reorganizes an active dispute about measurement error in LLM-generated exposure scores. Yin et al. (2026) show that occupational exposure scores can vary sharply across frontier models and that downstream labor-market coefficients can change sign with the rater model. Espinal Maya (2026) instead treats cross-model disagreement as classical attenuation that can be corrected using obviously related instrumental variables (ORIV; Gillen

[3]This diagnostic is computed on the 1,907 surveyed task statements, the subset for which the per-call rationale text is available.

et al., 2019). Both views can be right, but for different parts of the task space. When a task is self-contained, models answer a well-posed feasibility question and disagreement is close to noise around a common construct. When a task is institutionally co-dependent, the bare prompt omits the deployment environment, and models then infer that environment differently. The resulting score variation is part of the economic object rather than nuisance error.

Our contribution is twofold. First, we extend the task-based literature on AI exposure by incorporating institutional deployment constraints into the exposure measure itself (Brynjolfsson et al., 2018; Webb, 2020; Felten et al., 2021), and increasingly uses GPT and other LLMs to rate task-level feasibility (Eloundou et al., 2024; Eisfeldt et al., 2023; Gmyrek et al., 2023; Cazzaniga et al., 2024; Handa et al., 2025). Parshall and Lopez-Luzuriaga (2026) further separate cognitive complexity, physical deployment difficulty, and regulatory restrictions into distinct task characteristics. A complementary adoption literature studies the conditions under which technical capability enters production, including comparative advantage and user costs (Lindenlaub et al., 2026), integration into enterprise workflows (Yu et al., 2026), and managerial strategy, trust, and organizational complementarities (Makridis, 2026). Finance provides a useful setting in which to connect these literatures because it combines high technical exposure and rapid AI use with specific governance requirements (Makridis et al., 2026). We estimate, for the same task and model, how the share of human time that AI can eliminate changes when the task is placed inside a financial institution and evaluated under finance-specific governance. Holding the task, model checkpoint, score scale, and output schema fixed identifies a continuous institutional markdown from within-model revisions, while the eight-checkpoint design shows whether the adjustment appears across model families. In the NAICS 52 task universe, the institutional markdown averages one-fifth of technical feasibility and is concentrated in regulated and client-facing occupations.

Second, we contribute to the literature linking firm-level AI exposure to corporate outcomes and asset prices by separating the level of technical exposure from its composition. Existing work relates AI investment or exposure to firm growth, product innovation, market values, and systematic risk (Babina et al., 2024b; Eisfeldt et al., 2023; Babina et al., 2024a). Related asset-pricing studies infer beliefs about AI from market reactions to frontier-model releases (Andrews and Farboodi, 2025) or construct AI factors from realized model consumption (Borri et al., 2026). These approaches measure AI investment, exposure, news, or use as a scalar. We instead use pre-ChatGPT Revelio Labs occupation shares (Revelio Labs, 2026) to construct an ex ante, staffing-based measure and decompose technical exposure into deployable exposure and the institutional markdown. Because the two components sum to technical exposure in common units, a joint regression nests the scalar exposure specification: equal coefficients imply that total technical exposure is sufficient, while unequal coefficients imply that the composition of exposure contains additional information. Empirically, the return difference is concentrated in the markdown: firms above the median underperform by roughly 25 percentage points in market-adjusted cumulative abnormal returns, and the post-ChatGPT estimate lies below all 58 rolling pre-event estimates, while sorts on technical and deployable exposure produce little difference. High-deployability firms also exhibit higher profitability and lower employment after 2022, whereas high-markdown firms exhibit lower employment without a profitability gain; because the monthly Fama–MacBeth slopes and factor-adjusted spreads are imprecise, we interpret this as evidence that the decomposition contains information beyond total technical exposure, not as a causal estimate of AI repricing.

We structure the rest of the paper as follows. Section 2 describes the task universe, model-scoring panel, benchmarks, and current data limitations. Section 3 defines technical feasibility, institutional markdowns, and the prompt-ladder decomposition. Section 4 implements the measure

and reports the measurement results. Section 5 constructs firm-level exposure from Revelio Labs firm-by-occupation data and reports the cumulative-return cross-section, the Fama–MacBeth and factor-model tests, and the placebo evidence. Section 6 concludes.

2 Data and Measurement Inputs

This section describes the data needed to construct occupation-level finance exposure. The unit is an O*NET task performed inside the finance-and-insurance industry. The empirical implementation scores rungs A–C of the prompt ladder and measures the institutional markdown from bare feasibility to named finance governance regimes.

2.1 Task universe

We begin with O*NET task statements and retain occupations that appear in finance and insurance employment, defined by NAICS 52. The resulting production universe contains 2,199 task statements across 99 six-digit SOC occupations. Of these, 1,907 are Core statements that O*NET has surveyed for importance and frequency. The remaining 292 are statements O*NET lists for the occupation but has not rated. We admit rather than discard them, because discarding them would drop six occupations entirely, including financial and investment analysts and financial risk specialists, whose statements O*NET has not rated at all. An unrated statement inherits the median importance of the rated statements in its own occupation, so it is treated as a typical task of that job. Where an occupation has no rated statement, every statement receives equal weight. The rule degrades gracefully, in the sense that the median carries less information as the rated share of an occupation falls and the scheme converges on equal weighting without a threshold.

While we focus on financial services (NAICS 52). we retain SOC occupations that clear the finance-employment floor, including customer-service representatives, software developers, marketing specialists, compliance officers, accountants, insurance claims workers, tellers, underwriters, loan officers, and credit analysts. This industry-based definition captures the occupations that financial institutions actually employ, since deployability depends on the setting in which a task is performed as well as the occupation title. We then merge each occupation's O*NET task statements, task importance, task frequency, and job-zone measures. Before any model is queried, the task-level file therefore contains the task text, occupation and finance-industry employment weights, and pre-scoring controls including word count, O*NET importance, frequency, and job-zone preparation requirements. Occupation-level exposure aggregates task scores using O*NET task weights. Let k index tasks and $o(k)$ the SOC occupation to which task k belongs. For model m, vintage t, and prompt rung r, the task score is $s^F_{kmtr} \in [0, 1]$. The occupation score is

$$E^F_{omtr} = \sum_{k \in o} \omega_{ko} s^F_{kmtr}, \qquad \sum_{k \in o} \omega_{ko} = 1, \tag{1}$$

where ω_{ko} is the normalized O*NET task weight. The main tables report both equal-occupation averages and NAICS-52 employment-weighted averages, because the first describes the task measurement problem and the second describes the exposure of finance employment.

2.2 Model-scoring panel

The scoring panel covers eight frontier model checkpoints from two independently developed closed-model families: GPT and Claude. We use a common API router so that prompts, output schemas, temperature, logging, and run metadata are held fixed as much as possible across providers. Each

task–model–rung cell is scored five times at temperature 0.7 and summarized by the median score. The score is an integer between 0 and 100 that we rescale to the unit interval for estimation.

The panel records the model family, resolved model identifier, query date, prompt template, rung, repetition, score, and structured-output diagnostics. This logging serves the measurement design as well as replication. Closed models can change behind stable public names, and different model families use numerical scales differently. We thus rely on within-model differences across rungs and on rank-based agreement measures—not on unadjusted cross-model score levels.

2.3 Prompt rungs

The exposure panel scores three rungs. Rung A asks the model to score bare technical feasibility: how much task time a capable model could save if the task were considered in isolation. Rung B holds the task fixed but states that the task is performed inside a financial institution. Rung C adds named finance governance domains—model-risk governance, adverse-action and explainability, confidentiality, supervisory review and recordkeeping, and accountable human review—but does not supply task-specific codes. Section 3 defines the institutional markdown constructed from these rungs, and the constraint codebook in Appendix A defines the institutional dimensions that anchor its interpretation.

2.4 External benchmarks

We use the task ratings of Eloundou et al. (2024) as the main external benchmark for the feasibility rung. Their ratings and ours share the same O*NET task identifiers, so we can aggregate both measures over the identical task set and identical task weights. The benchmark measures what

rung A is designed to measure: technical feasibility, not finance deployability. A high correlation between rung A and the Eloundou score validates the feasibility rung. A weaker correlation between the markdown and the Eloundou score indicates that the institutional layer adds information not contained in the canonical exposure measure.

The second benchmark measures realized usage rather than rated feasibility. The Anthropic Economic Index (Handa et al., 2025) reports each O*NET task's share of global Claude conversations, separately for claude.ai chat traffic and first-party API traffic, together with conversation-level diagnostics on collaboration patterns and AI autonomy. We use the June 2026 release, which covers April and May 2026. The index publishes a task only when its usage clears privacy and volume thresholds, so between 650 and 747 of the 2,199 universe tasks appear in a given source and month, and 97 of the 99 occupations appear at the occupation level. A task absent from a release is unpublished rather than unused, a censoring feature that the validation design in Section 4.5 models explicitly. Task-level usage is economy-wide rather than specific to finance-industry workers, and it covers a single model family. Both features add usage variation that is unrelated to the finance-specific markdown and should attenuate the estimated associations.

Other adoption benchmarks—the Census Bureau's Business Trends and Outlook Survey, nationally representative worker-adoption surveys, and within-firm AI adoption measures—will serve later validation exercises. The present results simply establish that the measurement instrument produces a stable and interpretable finance markdown and that the resulting deployable measure predicts realized usage, before asking whether it predicts employment, wages, or returns.

Nevertheless, we note at least two limitations. First, the markdown should be read as the effect of finance framing and named governance regimes, not as a task-specific deployability adjustment. Second, the task universe is finance-industry employment based. It captures the occupations

financial firms employ, but it does not incorporate firm-specific staffing shares. Section 5 introduces Revelio Labs data for that aggregation step.

3 Measuring Deployable Exposure

Having described the data, we now define the measurement object. We separate a task scored as an isolated activity from the same task performed inside a financial institution. The first object is technical feasibility. The second object is deployability. The empirical problem is that standard exposure scores usually observe only the first object, even though the second is the object relevant for firm adoption and asset prices.

3.1 The target construct

Let k index tasks, $o(k)$ occupations, m models, t model vintages or scoring dates, and $r \in \{A, B, C\}$ prompt rungs. The superscript F denotes that the task is evaluated inside finance rather than in an abstract occupation.

Definition 1 (Finance deployment mode). The baseline deployment mode is the lowest-friction compliant human-in-the-loop architecture available to a financial institution. The AI system may draft, summarize, classify, retrieve, analyze, or recommend. A human remains accountable for customer-facing, trading, credit, insurance, advisory, financial-reporting, regulatory, or supervisory actions. The institution preserves the information required for review, reconstruction, escalation, and audit. The score does not assume full autonomous deployment unless the task can be completed autonomously without changing these constraints.

Definition 2 (Finance-industry task score). For task k, model m, vintage t, and rung r, $s^F_{kmtr} \in$

$[0, 1]$ is the model-elicited share of human task time that AI could eliminate while preserving the relevant quality standard under the information supplied at rung r. Operationally, the model reports an integer from 0 to 100, which we rescale to the unit interval.

The measurement objects are

$$\text{Tech}^F_{kmt} = s^F_{kmtA}, \qquad \text{Inst}^F_{kmt} = s^F_{kmtC}, \qquad \Delta^{F,AC}_{kmt} = \text{Tech}^F_{kmt} - \text{Inst}^F_{kmt}. \tag{2}$$

Tech^F is bare technical feasibility. Inst^F is the score after finance context and named governance regimes are supplied. $\Delta^{F,AC}$ is the institutional markdown. Rung C tells the model which finance regimes exist. The coded constraint vector defined below characterizes which of those regimes bind for a specific task.

We identify the markdown from within-model differences. Cross-model levels mix capability, calibration, and scale use: one model may treat 70 out of 100 as a conservative score while another uses it aggressively. The markdown compares a model to itself across two versions of the same task. It removes the most visible level confound and asks a narrower question: how much does the same model revise exposure when the finance deployment environment is made explicit?

3.2 Why finance changes the task boundary

Finance is useful for this exercise because many deployment constraints are external to the rater model. A task may sit inside a credit workflow, a broker-dealer communication process, an investment-advice process, an insurance-claims file, a bank model-risk inventory, or a privacy and information-security program. Each implies observable requirements for review, documentation, explanation, retention, validation, and human sign-off.

To see this, compare two high-feasibility tasks. Summarizing a public 10-K for an internal analyst can be performed with public inputs, low customer risk, and a light review layer. Drafting an adverse-action explanation for a denied consumer loan may require customer-specific nonpublic information, a reviewable decision file, legally meaningful reason codes, and an accountable human decision-maker. Both tasks involve language, summarization, and financial reasoning. Only the second is tightly co-dependent on the institution.

For the same reason, we do not define finance exposure by finance-titled occupations alone. A software developer, marketing specialist, customer-service representative, lawyer, accountant, or data scientist can perform a finance-specific task when the task is embedded in a bank, insurer, broker-dealer, investment adviser, or payments firm. The current implementation captures this by using occupations employed in NAICS 52 and asking the scoring instrument to evaluate the task inside the finance industry.

3.3 Institutional co-dependence

We characterize the institutional constraints on each task with a coded constraint vector,

$$h_k^F = \left(h_k^{MRM,F}, h_k^{AA,F}, h_k^{CONF,F}, h_k^{SUP,F}, h_k^{ACC,F}\right), \tag{3}$$

where each component takes values in $\{0, 1, 2\}$: not binding, partially binding, or strongly binding. The five dimensions correspond to model-risk governance, adverse-action and explainability obligations, confidentiality and data restrictions, supervisory review and recordkeeping, and auditability and human accountability. Appendix A gives the coding rules.

We summarize the vector with

$$H_k^F = \frac{1}{10} \sum_{j \in \{MRM, AA, CONF, SUP, ACC\}} h_k^{j,F}, \tag{4}$$

so that $H_k^F \in [0, 1]$. Equal weights are the baseline because the five dimensions capture different channels, not five noisy proxies for a single latent factor. The scalar score asks whether high-co-dependence tasks have larger markdowns. The component scores ask whether the correct institutional channel binds for the correct task family: adverse-action constraints for credit decisions, supervisory-review constraints for client communications, confidentiality constraints for customer-data tasks, model-risk governance for recurring analytic systems, and accountability constraints for tasks in which a human must remain responsible for the final action.

The current results do not rely on the independent H_k^F vector. The appendix reports a diagnostic proxy built from the rung-C model responses. Because the proxy is derived from the same model outputs used to score the constrained rung, we treat it only as a pipeline diagnostic. The main evidence is the size of the markdown, its occupational pattern, its relationship to technical feasibility, the Eloundou validation, and the cross-model agreement gradient across rungs.

3.4 The prompt ladder as a decomposition

The prompt ladder is a measurement device, not a claim that firms deploy AI in stages. The rungs hold the task, model, score scale, and output schema fixed while changing only the institutional information supplied to the model. Rung A asks for bare technical feasibility. Rung B adds that the task is performed inside a financial institution but gives no named regimes. Rung C names the finance governance domains.

The institutional markdown decomposes as

$$\begin{aligned}\Delta^{F,AC}_{kmt} &= s^{F}_{kmtA} - s^{F}_{kmtC} \\ &= \underbrace{\left(s^{F}_{kmtA} - s^{F}_{kmtB}\right)}_{\text{finance-context adjustment}} + \underbrace{\left(s^{F}_{kmtB} - s^{F}_{kmtC}\right)}_{\text{named-regime adjustment}} .\end{aligned} \tag{5}$$

The empirical results show that the first term is larger than the second: most of the markdown occurs when the task is placed inside finance, while the named regulatory layer refines the score rather than imposing a large additional penalty.

3.5 Interpretation and limits

Deployable exposure is not realized adoption. A task can be deployable and still not be adopted because integration costs are high, vendors are immature, legacy systems are brittle, or managers have not reorganized workflows around the tool. Conversely, firms may experiment with tasks that receive low deployable-exposure scores, but those deployments should require more review, narrower scope, or more human labor than bare technical scores imply.

The constraint vector is also not a legal opinion. It does not ask whether a specific AI system is legally permissible for a specific firm under a specific supervisory posture. It asks a measurement question: how much institutional scaffolding is required before AI capability can become usable production inside finance?

In principle, constrained exposure should not exceed technical exposure. In practice, elicited scores may occasionally violate this monotonicity restriction because LLM scoring is stochastic or because added context clarifies a task rather than constraining it. We retain signed markdowns in the main analysis, report nonnegative and proportional versions as robustness checks, and use

job-zone and word-count diagnostics to show that the measured markdown is not simply a proxy for task length or generic task complexity.

4 Implementing the Measurement: Results

Having defined the instrument, we now implement it. We report what rungs A–C support: the feasibility score, the finance-context and named-regime markdowns, the occupational distribution of the $A-C$ markdown, the cross-model agreement pattern, and a validation of deployable exposure against realized AI usage. We do not treat the model-generated co-dependence proxy as a main result, and we do not estimate labor-market regressions in this section. The return tests appear in Section 5, once the occupation-level measure is aggregated to firms.

4.1 Benchmarking the feasibility rung

We first ask whether rung A measures the same object as the existing task-exposure literature. This is a necessary diagnostic. If the bare feasibility score does not line up with a canonical feasibility measure, the institutional markdown could be an artifact of the scoring instrument rather than new information about deployability.

The feasibility rung performs well. We merge our 2,199 tasks to the task ratings in Eloundou et al. (2024). Of these, 2,194 tasks merge directly, covering 99.8 percent of the task universe. We then aggregate both measures over the identical task set and identical O*NET weights. Across the 99 occupations, our rung-A score correlates 0.73 with the Eloundou GPT-4 rating and 0.53 with the Eloundou human annotation. The corresponding task-level correlations are 0.56 and 0.39 (Appendix Figure A.6 shows the task-level scatter). Figure 1 shows the occupation-level

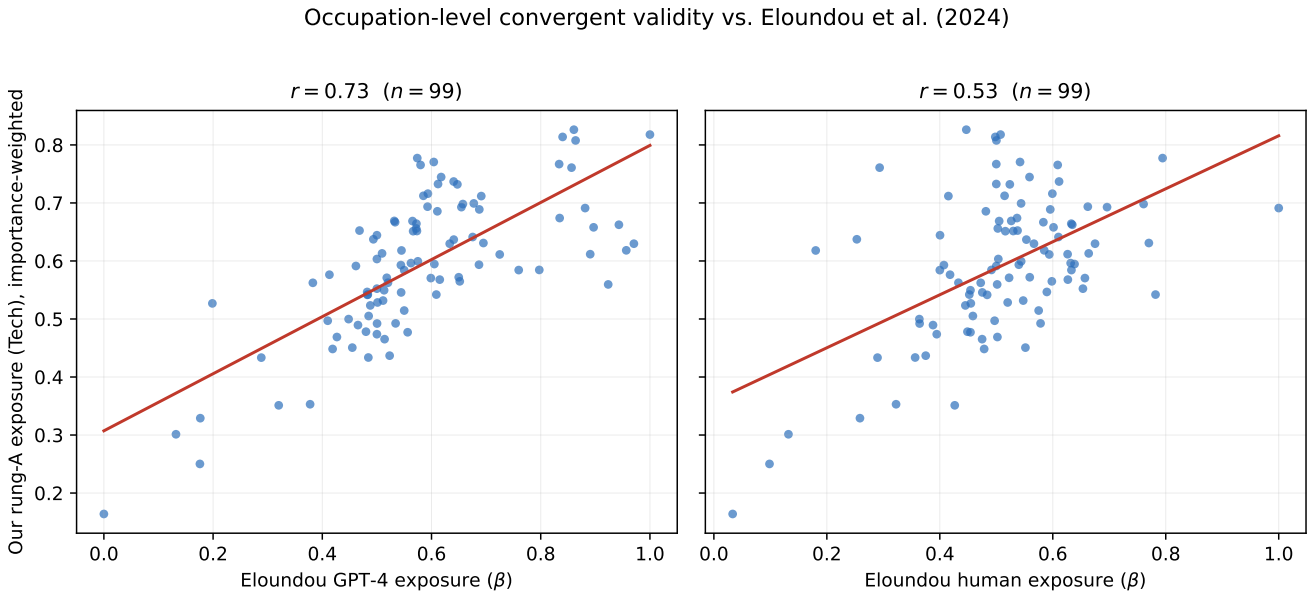


Figure 1: Occupation-level convergent validity against Eloundou et al. (2024). Each point is one of the 99 finance-and-insurance occupations. The vertical axis is our importance-weighted rung-A technical-feasibility score. The horizontal axis is the Eloundou β exposure aggregated over the identical tasks and weights, using the GPT-4 rating in the left panel and the human-annotator rating in the right panel. The 2,194 merged tasks cover 99.8 percent of the task universe.

relationship. The fit is tighter for GPT-4 than for the human annotation, which is useful for interpretation: model-based raters agree strongly about technical feasibility, but not perfectly.

The markdown correlates only 0.23 with the Eloundou GPT-4 rating at the occupation level and 0.21 at the task level. The weak correlation is the prediction of the construct, not a validation failure: rung A should agree with a feasibility benchmark, while the markdown should capture institutional content that the benchmark omits by design. The benchmark therefore validates the feasibility rung and shows that the institutional markdown is not a rescaled version of existing O*NET exposure. An occupation's markdown cannot be recovered from its position in the existing exposure rankings. Recovering it requires the institutional layer of the instrument.

4.2 The average markdown is large

Table 8 reports the unweighted exposure means by model and prompt rung. Every model assigns lower exposure once the task is placed inside finance and evaluated under named governance regimes. The raw $A - C$ markdown ranges from 1.2 points for GPT-4.1 to 17.2 points for Claude Sonnet 4. The result is not driven by a single model family: both GPT and Claude checkpoints mark down exposure once the institutional environment is introduced.

Table 9 reports model-specific importance-weighted estimates and the pooled employment-weighted estimate, with clustered confidence intervals. The pooled employment-weighted markdown is 12.0 points, with a 95% confidence interval from 9.6 to 14.8. The equal-occupation, importance-weighted estimate is 9.5 points. To gauge the magnitude, the mean rung-A feasibility score is 62 points, so finance context and named governance regimes remove roughly one-fifth of measured technical exposure for the average finance worker. All eight model-level markdowns are positive and statistically distinguishable from zero. The family gap is large in points—leaving out Claude models produces a 7.2-point markdown, while leaving out GPT models produces a 16.9-point markdown, a difference with a 95% confidence interval from 4.7 to 15.3 points—but the sign of the result does not depend on either family. Because the markdown is a within-model difference, the family gap is a substantive finding about how differently trained models internalize institutional constraints.

The decomposition in eq. (5) locates the markdown. The mean $A - B$ adjustment is 10.1 points, while the mean $B - C$ adjustment is 1.9 points. Most of the markdown occurs when the task is first embedded in finance, not when the prompt subsequently lists named regulatory domains. The models are not reacting to a list of statutes. They reprice the task as soon as they

learn the production environment. Figure 7 shows the same fact distributionally: the task-score density shifts left between rung A and the finance-governed rungs rather than changing only in a small tail of tasks. The markdown is a broad repricing of the finance task universe, not a penalty applied to a handful of obviously regulated tasks.

4.3 The markdown is concentrated in regulated, client-facing occupations

The average markdown masks substantial occupational heterogeneity. Table 10 lists the fifteen largest and fifteen smallest occupation-level markdowns, and Figure 5 plots the same contrast with confidence intervals. The largest markdowns appear in occupations where AI output is tied to customer decisions, sensitive data, regulated communications, or accountable review: loan interviewers and clerks, credit analysts, credit authorizers, new-account clerks, loan officers, insurance underwriters, credit counselors, tellers, securities-sales agents, customer-service representatives, bill and account collectors, insurance claims and policy processing clerks, and personal financial advisers. These occupations combine high technical feasibility with high institutional co-dependence, and they map directly onto the regulatory anchors in the codebook: adverse-action and explainability rules for the credit roles, supervisory-review and recordkeeping requirements for the sales and advice roles, and confidentiality obligations for the customer-data roles.

The smallest markdowns appear in a different part of the task space: public-relations specialists, retail salespersons, market-research analysts, health-education specialists, security guards, administrative-services managers, public-relations managers, graphic designers, meeting planners, software developers, web-interface designers, training managers, mail clerks, janitors, and mainte-

nance workers. Some of these occupations are technically exposed, but their tasks are less often tied to credit decisions, customer files, supervised financial advice, or accountable production decisions inside finance.

Figure 6 plots the markdown against rung-A technical feasibility. The fitted relationship is positive: occupations that look most exposed in isolation also lose the most measured exposure once the task is evaluated inside finance. The upper-right region of the figure is the population the title of the paper describes—capable but not deployable. The pattern has a direct implication for firm-level measurement, so the markdown enters the firm-level exposure measure in Section 5 separately from the feasibility score rather than being netted against it. A firm staffed with high-feasibility, high-markdown occupations is not exposed in the same way as a firm staffed with high-feasibility, low-markdown occupations, so a firm-level exposure measure should carry the markdown separately from the feasibility score rather than netting the two against each other.

We conduct two diagnostics. Appendix Table A.3 shows a small negative association between task word count and the markdown once model fixed effects and job-zone dummies are included (-0.044 points per word; wild-cluster $p = 0.064$). Longer task statements therefore do not mechanically receive larger markdowns. Appendix Figure A.3 shows that the markdown is roughly flat across O*NET job zones. The markdown is therefore not a proxy for task verbosity, education requirements, or generic skill intensity. Appendix Table A.4 shows that the pooled markdown remains positive under signed, nonnegative, and proportional definitions. It also remains positive when one developer family is dropped at a time and when adherence outliers are excluded.

4.4 Models agree more on feasibility than deployability

The cross-model evidence points in the same direction. Table 11 reports the mean pairwise Spearman rank correlation across the eight models at each rung. Agreement is highest under bare feasibility: 0.804 at rung A. It falls to 0.768 at rung B and 0.738 at rung C. Figure 8 shows the corresponding increase in cross-model dispersion as institutional context is added. Models concur more on what a task is technically exposed to than on what is deployable inside finance. The rank agreement also conceals a systematic vendor gap in levels, shown descriptively in Appendix Figure A.5 using importance-weighted occupation means. All but one of the 99 occupations receive a higher exposure score from the GPT family than from the Claude family at rung A, and every occupation does so at rungs B and C. The mean gap widens from 8.7 points at rung A to 13.6 points at rung B and 17.0 points at rung C, even though the cross-vendor Spearman rank correlation stays high at 0.93, 0.90, and 0.85. Vendors preserve the ordering of occupations but diverge on levels, the divergence grows with institutional context, and single-vendor exposure levels are therefore not interchangeable.

This result is distinct from the average markdown. The mean score can fall without changing rankings across models, and rankings can diverge without a large average fall. We observe both. The level result says finance context lowers measured exposure. The agreement result says finance context also makes exposure harder for models to rank consistently. Together they locate the measurement-error dispute described in the introduction: the omitted institutional layer produces both a mean adjustment and instability in the exposure construct, so cross-model disagreement in finance is informative about co-dependence rather than pure noise to be averaged away.

Several additional diagnostics remain. Appendix Table A.2 summarizes the model-generated

primary binding constraint at rung C. Accountability, confidentiality, and supervisory review dominate. Appendix Tables A.5, A.6, and A.7 use the same model-generated diagnostic proxy as a preliminary co-dependence measure. Those tables are not main evidence. Because the proxy is derived from the rung-C responses, it carries a mechanical circularity concern. We therefore report these estimates only as pipeline diagnostics and do not use them to validate the markdown.

4.5 Deployable exposure predicts realized AI usage

The evidence so far compares the instrument to a rated benchmark and examines its internal structure, but it does not establish that the markdown is related to actual AI use. If the markdown reflects institutional co-dependence rather than an artifact of institutional prompt language, tasks with larger markdowns should see less realized usage.

The Anthropic Economic Index allows this comparison at the task level, because it reports realized Claude usage for individual O*NET task statements (Handa et al., 2025). As Section 2 describes, the June 2026 release covers April and May 2026 and separates chat traffic, most of which reflects individual use, from first-party API traffic, most of which reflects production deployment. The index publishes a task only when usage clears privacy and volume thresholds. Between 650 and 747 of the 2,199 universe tasks appear in a given source and month, and a task that does not appear is unpublished rather than unused. We therefore estimate usage regressions on published tasks and model publication separately as an extensive margin.

Table 1 relates task-level usage to the two exposure measures. The dependent variable is the log usage share of a published task, the exposure regressors are cross-model task means in standard-deviation units, the two months are pooled with month fixed effects, and standard errors

are clustered on occupation. When the measures enter separately, both predict usage. When they enter together, a one-standard-deviation increase in deployable exposure is associated with 54 log points more chat usage and 75 log points more API usage, and the coefficient on technical exposure becomes negative in both sources, at −42 and −47 log points. Because $\text{Tech}^F = \text{Inst}^F + \Delta^{F,AC}$, the coefficient on technical exposure in the joint specification measures the association between the markdown and usage at a given level of deployable exposure. Tasks with the same deployable exposure are used less when a larger share of their measured feasibility does not survive the institutional layer. The component of exposure that the instrument removes is therefore not classical measurement error, since it predicts usage with the opposite sign from the component that remains. This error structure explains the disagreement between existing exposure measures.

Table 1: Deployable exposure and realized task-level AI usage

	(1) Chat	(2) Chat	(3) Chat	(4) API	(5) API	(6) API
Technical exposure Tech^F	0.120		-0.422**	0.271***		-0.465***
	(0.0775)		(0.164)	(0.0735)		(0.125)
Deployable exposure Inst^F		0.188***	0.538***		0.357***	0.752***
		(0.0662)	(0.156)		(0.0656)	(0.120)
Wild-cluster p (deployable)			0.002			0.000
Observations	1,446	1,446	1,446	1,393	1,393	1,393
R^2	0.005	0.014	0.023	0.025	0.049	0.061

Notes. The dependent variable is the log usage share of a published task in the Anthropic Economic Index, June 2026 release: the task's percentage of global Claude conversations in the source named in the column header, where Chat is claude.ai traffic and API is first-party API traffic. The sample pools April and May 2026 with month fixed effects and covers the finance-universe tasks that clear the index's publication thresholds. Published shares are rounded to two decimals, so the 133 published tasks with rounded-to-zero shares enter at half the rounding resolution rather than being dropped. Exposure regressors are cross-model task means over the eight-checkpoint panel in standard-deviation units, with rung C standing in for deployable exposure. Standard errors are clustered on occupation. The wild-cluster row reports the bootstrap p-value for the deployable-exposure coefficient. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Because the usage regressions condition on a task clearing the publication threshold, they select on the outcome. Table 2 addresses this selection by modeling publication directly. We expand the

universe to all 8,796 task–source–month cells and estimate a linear probability model, in percentage points, for whether a task is published. Technical exposure alone raises the publication probability by 13.0 points per standard deviation. When technical exposure is split into its two components, the deployable component raises the publication probability by 14.8 points per standard deviation, while a one-standard-deviation larger markdown lowers it by 3.3 points at a given level of technical exposure, with a wild-cluster p-value of 0.015. Relative to the baseline publication rate of 32 percent, one standard deviation of the markdown is associated with a roughly ten percent lower probability that a task shows any measurable usage.

Table 2: The markdown and the extensive margin of AI usage

	(1) Published	(2) Published	(3) Published
Technical exposure Tech^F	13.04***		14.83***
	(1.064)		(1.359)
Deployable exposure Inst^F		13.23***	
		(1.080)	
Markdown $\Delta^{F,AC}$			-3.307**
			(1.277)
Wild-cluster p (markdown)			0.015
Observations	8,796	8,796	8,796
R^2	0.079	0.081	0.083

Notes. The dependent variable equals 100 if the task is published in the Anthropic Economic Index in the given source and month and 0 otherwise. The sample is the full task universe expanded over the four source–month cells (2,199 tasks, two sources, two months), with cell fixed effects. A task absent from the index is unpublished rather than unused, which is the censoring this margin models. Exposure regressors are cross-model task means in standard-deviation units, with rung C standing in for deployable exposure. Standard errors are clustered on occupation. The wild-cluster row reports the bootstrap p-value for the markdown coefficient. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The manner in which high-markdown tasks are used is consistent with the constraints that generate the markdown. Appendix Table A.9 regresses the index's conversation diagnostics on the markdown, conditional on technical exposure. A one-standard-deviation larger markdown is

associated with a 1.7 percentage point higher share of conversations in the augmentation collaboration mode, a 0.4 percentage point higher share of validation-style interactions, and a 0.10 point lower assessed AI autonomy on a five-point scale, while technical exposure enters with the opposite sign in each column. Usage of high-markdown tasks involves more human review and less delegation, the pattern one would expect if the accountability, supervisory-review, and confidentiality constraints in the codebook were binding, rather than the uniform reduction in usage one would expect if the markdown reflected missing capability. Appendix Table A.10 examines the allocation of usage across the two sources. The ratio of chat usage to API usage does not vary with the markdown, so high-markdown tasks are not differentially absent from production traffic in this period. The gap between the API and chat automation shares does widen with the markdown, by 3.5 percentage points per standard deviation. In these two months the constraints are visible in the supervision of usage rather than in the volume of production usage.

The relationship also holds at the occupation level, where the labor-market outcomes of interest are measured. Figure 2 plots occupation-level usage against occupation deployable exposure for the published finance-and-insurance occupations, and Appendix Table A.11 reports the corresponding regressions with a log-employment control and source-by-month fixed effects. Occupation usage rises by 0.56 log points per standard deviation of deployable exposure, technical exposure alone is not a statistically significant predictor, and the occupation markdown enters at -1.18 log points per standard deviation. Occupations with large markdowns use AI less than their technical feasibility alone would suggest—a useful validation and consistent with theory.

Several qualifications apply. The estimates are associations between exposure scores and usage shares, not causal effects of exposure on adoption. The usage data cover a single model family over two months. Task-level usage is measured economy-wide rather than within the finance industry,

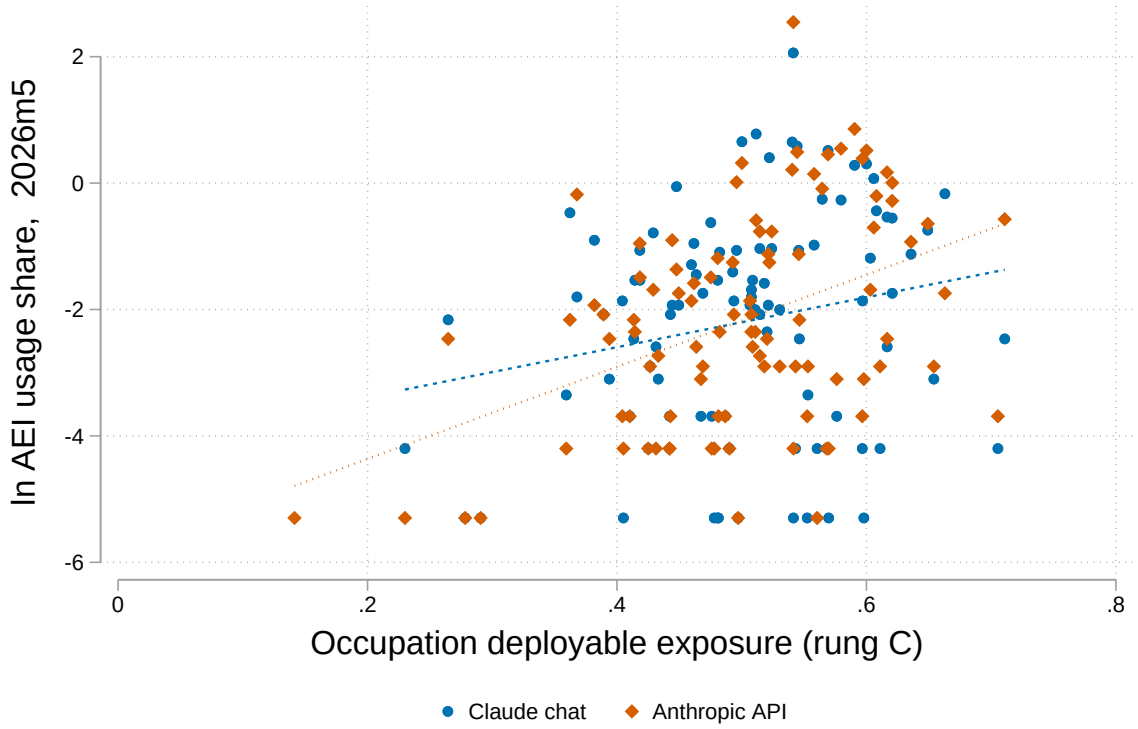


Figure 2: Occupation-level AI usage and deployable exposure. Each point is a finance-and-insurance occupation published in the Anthropic Economic Index in May 2026, with chat and first-party API traffic plotted separately. The vertical axis is the log of the occupation's share of global Claude usage in that source. The horizontal axis is the occupation's deployable exposure: the importance-weighted rung-C score over the occupation's tasks, averaged over the eight model checkpoints. Lines are OLS fits. Appendix Table A.11 reports the regression counterpart with both months, a log-employment control, and occupation-clustered inference. 97 of the 99 universe occupations clear the index's publication thresholds.

which if anything attenuates the association between a finance-specific markdown and usage. Finally, because half of the scoring panel consists of Claude checkpoints and the outcome measures Claude traffic, we use cross-model means pooled over both developer families as regressors, and rung C stands in for deployable exposure as in the rest of the wave-1 results.

4.6 The feasibility rung is not institution-free

The rationales that accompany the scores show why the bare-feasibility rung already carries institutional content. Every scoring call returns a short free-text rationale alongside the numeric score, and a keyword lexicon over the five governance regimes tags whether a rationale invokes finance-institutional language.[4] At rung A the prompt never names an industry, yet on average 17 percent of rationales per task invoke finance-institutional language such as confidentiality, supervisory review, or accountability, and 86 percent of tasks draw at least one such rationale across the eight models. The models read the deployment environment off the task wording itself. This spontaneous institutional salience predicts the markdown: tasks in the top quartile of the rung-A institutional-language share take an average markdown of 10.5 points, against 6.8 points in the bottom quartile, with a task-level correlation of 0.17. The pattern is direct evidence for the mechanism described in the introduction—when the bare prompt omits the deployment environment, models infer it from the task text. It also implies that rung A already embeds part of the institutional signal, so the measured $A - C$ markdown is a lower bound on the wedge between context-free feasibility and deployable exposure. The tags are lexicon-based and descriptive, and no clustered inference stands behind these particular numbers. This subsection is computed on

[4] The lexicon is calibrated against the structured constraint label that models emit at rung C, where the lexicon's top theme matches the model-emitted label for 68 percent of rationales.

the 1,907 surveyed task statements rather than on the full universe.

4.7 Model vintage evidence is descriptive

Appendix Table A.8 and Appendix Figure A.4 ask whether scores trend across model release dates. We keep these exhibits in the appendix because the comparison is underpowered and partly contaminated by the rubric's indexical "current frontier" language. The within-model markdown is more reliable. In the present data, no rung level or pooled markdown trends significantly with release date once task fixed effects and family structure are accounted for. This is a caution against overinterpreting cross-vintage levels, not as evidence that capability is flat over time.

4.8 What the current results establish

The current results establish five facts. First, the bare feasibility score validates against an external O*NET exposure benchmark, with an occupation-level correlation of 0.73. Second, finance context generates a positive within-model markdown of 12.0 employment-weighted points—one-fifth of measured feasibility—across every model in the panel. Third, the markdown is largest in the regulated, client-facing occupations where deployment should be most institutionally co-dependent, and it is not an artifact of task length or job-zone complexity. Fourth, cross-model agreement falls as institutional context is added. Fifth, deployable exposure predicts realized AI usage at the task and occupation level, while the component of technical exposure that the markdown removes is associated with less usage and less autonomous usage. The first four are measurement results. The fifth connects the measure to realized adoption.

5 Firm Exposure and Asset Prices

Next, we map the occupation-level measures to firms using pre-ChatGPT staffing composition. We construct firm-level technical exposure, deployable exposure, and the institutional markdown from Revelio Labs which covers the employment compositions (in SOC codes) for publicly traded firms, then examine how these measures relate to post-ChatGPT returns. We assess the timing and robustness of the return differences before turning to employment and profitability. Throughout, rung C measures deployable exposure and Δ^{AC} measures the institutional markdown.

5.1 Firm-by-occupation exposure

Rung C measures deployable exposure and Δ^{AC} measures the institutional markdown. We use Revelio Labs data to map occupation-level exposure into firm-level exposure. Revelio aggregates public professional profiles, job postings, and related workforce records into standardized firm, occupation, seniority, geography, and worker-transition data (Revelio Labs, 2026). The relevant input for this paper is the firm-by-occupation employment share: for public firm i, occupation o, and date t, let $\pi_{io,t}$ denote the share of the firm's workforce employed in occupation o. The firm-level exposure variables are weighted averages of the occupation-level measures:

$$\text{Tech}_{it}^{F} = \sum_{o} \pi_{io,t_0} \text{Tech}_{o}^{F}, \qquad \text{Inst}_{it}^{F} = \sum_{o} \pi_{io,t_0} \text{Inst}_{o}^{F}, \qquad \Delta_{it}^{F,AC} = \sum_{o} \pi_{io,t_0} \Delta_{o}^{F,AC}, \tag{6}$$

where t_0 is a pre-event staffing date. We fix the weights before the return window so that post-event hiring, layoffs, or occupational reclassification cannot respond mechanically to the same

AI shock whose price effect we measure. In the implementation, t_0 is the average over January through October 2022, the last ten months before the release of ChatGPT on November 30, 2022.

The firm-level variables have a direct interpretation. Tech_{it}^{F} measures how much of the firm's existing labor task bundle is technically exposed to AI. Inst_{it}^{F} measures the same bundle after finance context and named governance regimes are imposed. $\Delta_{it}^{F,AC}$ measures the share of technical exposure that disappears once those regimes bind. The identity $\text{Tech}_{it}^{F} = \text{Inst}_{it}^{F} + \Delta_{it}^{F,AC}$ means that return regressions cannot include all three variables without collinearity. We therefore parameterize the main specifications in Inst_{it}^{F} and $\Delta_{it}^{F,AC}$, which split technical exposure into its deployable and non-deployable components. Although Tech^{F} and Inst^{F} correlate at 0.70 across sample firms, the two components are not interchangeable in the cross-section, since sorting firms on the markdown produces a different high group than sorting on either level.

This design follows the logic of Eisfeldt et al. (2023), who aggregate occupation-level generative-AI exposure to publicly traded firms using firm occupational structure and study market reactions to AI news. Our extension is to decompose firm exposure into the part that survives finance deployment constraints and the part that those constraints mark down. The decomposition converts the measurement result of Section 4 into a pricing test: two firms with identical technical exposure can differ substantially in deployable exposure, and only the deployable component should move cash-flow expectations if institutions constrain what AI output can enter production.

The estimation sample starts from the 823 publicly listed U.S. finance and insurance firms with Revelio staffing coverage. Matching by CUSIP to CRSP common shares over 2015–2025 and requiring that at least half of a firm's positions map to a scored occupation and that the firm carries at least 25 weighted positions in the pre-event window leaves 576 firms. Monthly abnormal returns are computed relative to the CRSP value-weighted market return and winsorized at the first and

ninety-ninth percentiles within month. Accounting controls come from Compustat, aligned to the fiscal year preceding each return month.

5.2 Cumulative abnormal returns after ChatGPT

The sharpest capability event in the sample is the release of ChatGPT, which moved generative AI from a research artifact to a deployable input in a single news cycle. The first exercise therefore compares market-adjusted cumulative abnormal returns across exposure groups over the three years following the release. For each firm we sum the raw monthly market-adjusted return, $r_{it} - r_{mt}$, from December 2022 through December 2025,

$$CAR_i = \sum_{t=2022\text{m}12}^{2025\text{m}12} (r_{it} - r_{mt}) , \tag{7}$$

winsorize the resulting firm-level cross-section at the first and ninety-ninth percentiles, and regress it on an indicator for above-median exposure with finance-subsector fixed effects and standard errors clustered on subsector. The measure is additive rather than compounded, so decompositions by subperiod are exact and the robustness columns reported below reconcile with the baseline. The winsorization applies once, to the cumulated cross-section, so the long-horizon outcome is not a sum of monthly winsorized returns. The monthly regressions later in the section instead winsorize market-adjusted returns within month.

The headline sample is balanced: of the 560 firms with post-window returns, it keeps the 501 present in every month of the window, so that differences in cumulation horizon cannot masquerade as differences in performance. Twelve of these firms are the only sample firm in their subsector and drop out of the fixed-effects estimation, which is why the table reports 489

observations without controls and 404 once firms with incomplete pre-event accounts are excluded. Balance also conditions on survival: firms that delisted during the window, including the banks that failed in the spring of 2023, are absent by construction. Exit from the window is uncorrelated with the exposure sorts—regressing the balance indicator on the high-markdown indicator yields a coefficient of 0.01 with a t-statistic of 0.5—so survivorship is unlikely to generate the cross-sectional gaps, though the estimates remain effects among surviving firms.

Two features of the specification keep the cross-section free of post-event information. Controls are frozen at their October 2022 values because later values of size, profitability, or revenue growth are themselves outcomes of the shock being studied. The subsector that the fixed effects absorb and the clustering uses is likewise each firm's three-digit SIC as of October 2022, since collapsing the panel to one observation per firm would otherwise partition firms on a December 2025 industry classification. Eighty of the 576 sample firms change three-digit SIC between the two dates. Freezing the classification barely moves the estimates, taking the high-markdown coefficient from -0.255 to -0.250 without controls and from -0.222 to -0.205 with them, but it closes the last channel through which a post-event variable enters the design.

Inference uses the wild-cluster bootstrap. The estimation has twenty-four pre-event subsector clusters without controls and twenty-three with them, few enough that the asymptotic cluster-robust t over-rejects, and in this sample the two procedures disagree materially. We therefore report bootstrap p-values from 9,999 replications as a visible row in every long-horizon table and omit significance stars, so that no coefficient is displayed at a confidence level the small-cluster procedure does not support. Conventional clustered standard errors remain in parentheses for scale. Every significance statement about these cross-sections refers to the bootstrap p-values.

Table 3 reports the result. Firms in the top half of the markdown distribution underperform

the bottom half by 25.0 percentage points over the window, with a wild-cluster p-value of 0.05, and by 20.5 points once the frozen controls enter, with a p-value of 0.11. The movement is not an artifact of the accompanying sample change, since on the 404 firms with complete pre-event accounts the no-controls estimate is 24.3 points. Firms sorted on technical exposure alone show no gap, at -4.7 points with a p-value of 0.76, and firms sorted on deployable exposure show none either, at 7.0 points with a p-value of 0.28. The pattern is the one the construct predicts. What separates firms in the cross-section of cumulative returns is not AI exposure as such but the share of that exposure which governance constraints keep out of production. The economic magnitude is substantial. A market-adjusted cumulative shortfall of twenty-five percentage points over three years, concentrated among firms whose task bundles are capable but not deployable, is comparable to the cross-sectional dispersion generated by the interest-rate repricing of the same period.

The three median splits are separate regressions and are not, by themselves, a decomposition of technical exposure. Appendix Table A.13 runs the decomposition directly by entering deployable exposure and the markdown together in continuous form. Because $\text{Tech}^F = \text{Inst}^F + \Delta^{F,AC}$, markets that price bare feasibility would deliver equal coefficients on the two components, so equality is a nesting test of the standard feasibility-based specification. The test requires common units. Standardizing the two components separately rescales them by different factors and breaks the identity, after which coefficient equality no longer represents the restriction, so we divide all three components by a single factor, the standard deviation of technical exposure in the balanced cross-section, and read the coefficients as cumulative abnormal return per one technical-exposure standard deviation of the component. Technical exposure alone earns essentially nothing, at -0.001 with a p-value of 0.98. Splitting it, deployable exposure earns 0.151 and the markdown -0.210, and equality of the two is rejected with a wild-cluster p-value of 0.04. Once the frozen

controls enter, the two coefficients fall to 0.009 and −0.096 and equality can no longer be rejected, at a p-value of 0.60. The continuous specification therefore supports the same reading as the splits without controls and is uninformative about the restriction with them, which is why the median-split cross-section carries the statistical weight of the result.

We use the cumulative abnormal return as the headline long-horizon outcome because it is additive, permits exact decomposition by subperiod, and is far less sensitive to the right tail of compounded returns. Left unwinsorized, the high-markdown coefficient is −0.216, and it moves only to −0.182 when the cross-section is trimmed at the fifth and ninety-fifth percentiles, whereas the buy-and-hold abnormal return runs from −0.552 to −0.236 over the same range of trims. Appendix Table A.14 reports the two conventional compounded alternatives on the same balanced sample and specification. The negative high-markdown differential appears under both, at −33.8 percentage points in buy-and-hold abnormal returns and −20.0 points in relative wealth, with a wild-cluster p-value of 0.18 in each case. These are not independent tests of the same hypothesis. On a common balanced window the market wealth path is a single constant, so buy-and-hold abnormal returns and relative wealth are affine transformations of the same terminal firm wealth, which is why they share a p-value. They are reported so that the choice of the cumulative abnormal return reads as a methodological decision rather than an outcome-driven one.

Figure 3 plots cumulative abnormal returns for firms above and below the median of deployable exposure, without covariate or subsector adjustment. The figure is simply descriptive. The raw gap it displays arises largely from differences across subsectors—depository institutions sit mostly in the low-deployability group and performed worst over the window—and the subsector fixed effects in Table 3 absorb precisely this variation, which is why the table attributes no statistically distinguishable within-subsector cumulative gap to the deployable sort. The markdown sort, which

Table 3: Market-adjusted cumulative abnormal returns, December 2022 to December 2025

	(1)	(2)	(3)	(4)	(5)	(6)
High technical exposure	-0.047	-0.056				
	(0.077)	(0.052)				
High deployable exposure			0.070	-0.026		
			(0.064)	(0.070)		
High markdown					-0.250	-0.205
					(0.053)	(0.057)
ln(Assets)		-0.000		0.003		-0.011
		(0.024)		(0.022)		(0.022)
EBITDA/Assets		-2.333		-2.281		-2.382
		(1.329)		(1.296)		(1.331)
Revenue growth		-0.084		-0.080		-0.094
		(0.137)		(0.137)		(0.143)
Cash/Assets		1.022		1.049		0.916
		(0.399)		(0.398)		(0.409)
Prof. vol.		5.600		5.716		5.161
		(2.385)		(2.455)		(2.328)
Wild-cluster p-value	0.757	0.402	0.275	0.726	0.048	0.108
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	No	Yes	No	Yes	No	Yes
N	489	404	489	404	489	404
R-squared	0.17	0.29	0.17	0.29	0.18	0.30

Notes. The dependent variable is the firm's market-adjusted cumulative abnormal return, $CAR_i = \sum_t (r_{it} - r_{mt})$, summed over December 2022 through December 2025 and winsorized once, as a cross-section, at the 1st and 99th percentiles. The sample is the balanced panel of firms present in every month of the window. "High" indicators mark firms above the cross-sectional median of the pre-ChatGPT exposure variable named in the row, computed from January–October 2022 staffing weights. Medians are computed across the 576 firms in the exposure sample. Firm controls are log assets, EBITDA over assets, revenue growth, cash over assets, and five-year profitability volatility, frozen at their October 2022 fiscal-aligned values. All columns include pre-event finance-subsector fixed effects, defined by each firm's three-digit SIC as of October 2022 rather than the code carried by its final observation. Parentheses report standard errors clustered by the same pre-event subsector and are shown for scale. The row labeled "Wild-cluster p-value" reports two-sided wild-cluster bootstrap p-values based on 9,999 replications, clustered by the same subsector. Because the number of subsector clusters is small, at twenty-four without controls and twenty-three with them, statistical significance in the text is assessed using the wild-cluster p-values. No significance stars are reported.

does carry a within-subsector gap, is the object of the regression evidence.

Figure 4 disciplines the timing with an event-study specification: monthly abnormal returns are regressed on the high-exposure indicators interacted with calendar half-year bins, with firm and month fixed effects, so each coefficient measures the differential return relative to the first half of 2022, separately for the deployable and markdown sorts. For the deployable sort the differential is modest in the release half-year, emerges sharply in the first half of 2023 at close to two percentage points per month, and runs between one and one-and-a-half percentage points per month through the first half of 2025 before fading at the end of the sample. For the markdown sort the differential is largest in the first half of 2023, at nearly seven percentage points per month, and recurs at three to four points in the first halves of 2024 and 2025. Appendix Table A.12 aggregates the same interactions into three post-period bins against the full 2018–2022 pre-period and never rejects the hypothesis that the 2025 differential equals the 2023 differential, with p-values between 0.21 and 0.55 across specifications. The continued divergence of the cumulative series after 2024 therefore reflects the accumulation of a persistent monthly differential rather than a discrete repricing when agentic-AI capabilities arrived late in 2024. The timing evidence separates deployability repricing from the agentic-capability news cycle, although it cannot distinguish a single early-2023 repricing whose differential persists from a gradual repricing at a stable monthly rate.

Two features of the timing deserve emphasis. First, the largest single differential coincides with the March 2023 banking turmoil, which raises the concern that the result reflects bank distress rather than AI exposure. Section 5.4 addresses this directly and finds that the episode accounts for about a fifth of the full-window differential rather than for all of it. Second, the event-study coefficients are statistically flat over 2021 through the release, so the post-period differential does not continue a pre-existing trend over that horizon. The reference period itself sits at the bottom

of the 2022 rate-shock selloff for high-deployability firms, and coefficients for 2020 are nonzero, so the parallel-trends statement is local to the two years around the event rather than global.

5.3 Monthly pricing and factor exposure

The cumulative result raises the question of whether the differential is detectable at monthly frequency and whether it survives adjustment for standard return factors. Table 4 addresses the first question with Fama and MacBeth (1973) regressions. In every post-period month, the cross-section of abnormal returns is regressed on the standardized pre-ChatGPT exposure measures, the lagged firm controls, and finance-subsector dummies, and the table reports the time-series averages of the monthly slopes with Newey-West standard errors. The point estimates line up with the cumulative evidence. Technical exposure earns essentially nothing, at four basis points per month per standard deviation. Deployable exposure earns a positive but imprecise 10 basis points. The markdown earns -12 basis points, or -14 in the joint specification, which sums to roughly 5 percentage points per standard deviation over the thirty-seven-month window.

Because the median split separates groups differing by 1.7 standard deviations of the markdown, that slope implies a group gap near 9 points, well below the 25.0 points the split delivers, a discrepancy consistent with a cross-sectional relationship that is concentrated rather than linear in exposure. No monthly exposure slope is statistically distinguishable from zero. In the pre-period placebo of the final column the deployable slope falls to a fifth of its post-period value, but the markdown slope is of comparable magnitude and of the opposite sign, so the monthly cross-section on its own does not separate the two periods.

The portfolio version of the same comparison expresses the differential in more familiar units.

Table 4: Fama-MacBeth regressions of monthly abnormal returns on exposure

	(1)	(2)	(3)	(4)	(5)
Technical exposure (std.)	0.0004				
	(0.0013)				
Deployable exposure (std.)		0.0010		0.0011	0.0002
		(0.0012)		(0.0012)	(0.0013)
Markdown (std.)			-0.0012	-0.0014	0.0014
			(0.0015)	(0.0016)	(0.0018)
ln(Assets)	0.0007	0.0006	0.0005	0.0004	0.0008
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
EBITDA/Assets	-0.0336	-0.0334	-0.0366	-0.0357	0.0274
	(0.0287)	(0.0290)	(0.0298)	(0.0298)	(0.0241)
Revenue growth	0.0001	0.0003	-0.0002	-0.0001	-0.0010
	(0.0045)	(0.0044)	(0.0047)	(0.0046)	(0.0038)
Cash/Assets	0.0298**	0.0291**	0.0276**	0.0270**	0.0027
	(0.0129)	(0.0126)	(0.0121)	(0.0118)	(0.0120)
Prof. vol.	0.0865	0.0838	0.0781	0.0733	-0.0806
	(0.0922)	(0.0906)	(0.0882)	(0.0869)	(0.0726)
Sample	Post	Post	Post	Post	Pre
Subsector dummies	Yes	Yes	Yes	Yes	Yes
Months	37	37	37	37	59
Avg. firms/month	459	459	459	459	393
Avg. R^2	0.24	0.25	0.24	0.25	0.27

Notes. Fama-MacBeth estimates. In every month, the cross-section of winsorized market-adjusted returns is regressed on the standardized pre-ChatGPT exposure variables named in the rows, the lagged firm controls, and pre-event finance-subsector dummies defined by each firm's three-digit SIC as of October 2022. Reported coefficients are time-series averages of the monthly slopes. Standard errors in parentheses apply the Newey-West correction with three lags to the slope series, and t-statistics use $T-1$ degrees of freedom. Columns (1)–(4) use the post-period months, December 2022 through December 2025. Column (5) estimates the joint specification over the pre-period months with nonmissing controls, January 2018 through November 2022. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

An equal-weighted portfolio long high-deployability firms and short low-deployability firms earns 50 basis points per month after the release with a Newey-West t-statistic of 1.7, and the corresponding markdown spread loses 85 basis points per month with a t-statistic of 1.3. With only 37 post-event months, differentials of this size are estimated imprecisely at monthly frequency even though the corresponding market-adjusted cumulative return difference is economically large.

Table 5 evaluates the spreads against the five Fama and French (2015) factors plus momentum (Fama and French, 1993). Two findings emerge. First, the markdown spread loads significantly on the size and value factors. High-markdown firms are small value stocks, and part of their raw underperformance reflects factor structure rather than AI-specific repricing. The post-period alphas retain the predicted signs, at 47 basis points per month for the deployable spread and -30 basis points for the markdown spread, but neither is statistically distinguishable from zero. Second, the deployable-spread alpha before the release is five basis points, an order of magnitude below the post-period point estimate, so that estimate is not a mechanical artifact of the portfolio construction. The pre-period markdown and technical spreads do carry marginally significant positive alphas of roughly 30 basis points per month, which is the factor-structure behavior the placebo decomposition in Section 5.4 examines. The factor-adjusted monthly evidence is therefore directionally consistent with the cumulative result without adding independent statistical support, and the paper's pricing claim rests on the cumulative cross-section rather than on a monthly alpha.

Wild-cluster inference addresses dependence across firms within finance subsectors and the small number of subsector clusters, conditional on the realized December 2022–December 2025 period. It does not capture uncertainty across alternative event windows: all firms are observed during the same interest-rate cycle, banking turmoil, and sequence of AI releases. The monthly Fama–MacBeth and factor-spread estimates use time-series variation but are imprecise over only

37 months. We therefore compare the post-ChatGPT cumulative coefficient with estimates from the identical specification applied to 58 pre-ChatGPT windows of the same length. This exercise shows whether the post-event estimate is unusual relative to the historical distribution generated by the design, although it does not establish that ChatGPT caused the return differential.

Table 5: Hi-minus-lo exposure spread portfolios and Fama-French factors

	Deployable	Deployable	Markdown	Markdown	Technical	Technical
Alpha (monthly)	0.005	0.001	-0.003	0.003*	-0.002	0.003*
	(0.003)	(0.001)	(0.005)	(0.002)	(0.004)	(0.002)
mktrf	0.121	0.076***	-0.140	-0.363***	0.041	-0.224***
	(0.083)	(0.023)	(0.130)	(0.052)	(0.137)	(0.035)
smb	0.178*	-0.119***	0.300**	0.348***	0.301**	0.297***
	(0.093)	(0.040)	(0.119)	(0.083)	(0.117)	(0.066)
hml	-0.184**	0.076*	0.500**	0.382***	0.429**	0.321***
	(0.086)	(0.045)	(0.187)	(0.063)	(0.157)	(0.055)
rmw	-0.162	-0.117**	-0.022	0.245**	-0.043	0.156**
	(0.105)	(0.045)	(0.204)	(0.093)	(0.230)	(0.066)
cma	0.095	-0.198***	0.135	-0.275**	0.133	-0.274***
	(0.115)	(0.070)	(0.206)	(0.110)	(0.203)	(0.088)
umd	0.052	0.028	0.103	0.044	0.014	0.057
	(0.062)	(0.033)	(0.221)	(0.070)	(0.199)	(0.048)
Sample	Post	Pre	Post	Pre	Post	Pre
N	37	95	37	95	37	95

Notes. Each column is a time-series regression of a monthly hi-minus-lo spread portfolio return on the Fama-French five factors and momentum. Spreads are equal-weighted returns of firms above minus below the cross-sectional median of the pre-ChatGPT exposure variable named in the column pair. "Post" columns use December 2022 through December 2025 and "Pre" columns use January 2015 through November 2022. Newey-West standard errors with three lags in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.4 Alternative explanations

The return analysis uses three safeguards by construction. Firm exposure is computed from staffing weights frozen before the event. The main specifications include pre-event finance-subsector fixed

effects, so identification comes from within-subsector occupational structure rather than from comparing banks, insurers, asset managers, and payments firms with different business models, and the subsector classification itself is measured before the event. Controls in the cumulative cross-section are frozen at pre-event values. Beyond these, Table 6 confronts the two alternative explanations that the timing and the sample composition make most pressing.

Table 6: The markdown penalty: banking-crisis confound and timing placebos

	(1)	(2)	(3)	(4)	(5)	(6)
High markdown	-0.250	-0.245	-0.196	-0.052	-0.047	-0.179
	(0.053)	(0.096)	(0.043)	(0.022)	(0.085)	(0.081)
Window	Post window	Ex-depositories	Ex-2023m3-m5	Crisis window only	Placebo 2016-18	Placebo 2019-21
Wild-cluster p-value	0.048	0.033	0.017	0.137	0.540	0.407
N	489	257	489	505	372	419
R-squared	0.18	0.13	0.13	0.31	0.12	0.18

Notes. Each column regresses a balanced-window market-adjusted cumulative abnormal return on the high-markdown indicator with pre-event finance-subsector fixed effects, defined by each firm's three-digit SIC as of October 2022, and standard errors clustered on the same subsector. Column (1) is the baseline December 2022–December 2025 window. Column (2) drops every firm whose pre-event two-digit SIC is 60, that is, all depository institutions. Column (3) excludes the March–May 2023 months from the sum. Column (4) sums over the March–May 2023 crisis window only. Columns (5) and (6) apply the identical design to placebo windows ending before the release of ChatGPT. Each column is balanced and winsorized within its own window and subsample, so magnitudes are not strictly comparable across columns. Parentheses report clustered standard errors for scale. The row labeled "Wild-cluster p-value" reports two-sided wild-cluster bootstrap p-values based on 9,999 replications and is the basis for every significance statement in the text. No significance stars are reported.

The first alternative is that the markdown penalty reflects the 2023 regional-banking crisis rather than AI exposure. High-markdown occupations concentrate in deposit-taking institutions, and the differential emerges in the crisis half-year. Three exercises bound how much of the result the episode can account for. Dropping every firm whose pre-event two-digit SIC identifies it as a depository institution leaves 257 non-bank firms in the estimation and a markdown penalty of 24.5 percentage points with a wild-cluster p-value of 0.03, close to the baseline. Excising the March–May 2023 months from the sum leaves the penalty at 19.6 points with a p-value of 0.02. Summing over the crisis window alone yields 5.2 points, so the episode contributes about a fifth of the full-window differential. Because the outcome is additive, these two columns account for

all but two-tenths of a point of the baseline, the small residual reflecting that balancing over three months admits sixteen firms the thirty-seven-month window excludes and that each column is winsorized within its own sample. The defensible statement is that the banking turmoil does not account for the whole of the differential, not that it is irrelevant to it. The penalty survives excluding depository institutions, but that subsample holds only 56 high-markdown firms across twenty-one subsectors, so we do not claim that the differential is present separately within each.

The second alternative is that high-markdown firms perform poorly in all periods, so that any window would produce the same gap. The 2016–2018 placebo window is close to zero, with a differential of -4.7 percentage points and a wild-cluster p-value of 0.54. The 2019–2021 window requires more care. Its point estimate of -17.9 points is economically sizable and not statistically distinguishable from the baseline, so cumulative returns over a single hand-picked pre-period window cannot separate the two periods. Decomposing the placebo window shows that it differs from the post-ChatGPT window in origin and factor structure. The placebo gap does not arise in the acute COVID crash, which contributes $+0.4$ points over February–April 2020, and excising those months moves the gap from -17.9 to -19.2 points. The gap is concentrated in the growth-led 2020 recovery, as the factor loadings of Table 5 predict for small value stocks, and the factor model absorbs it fully, with a markdown-spread alpha of $+20$ basis points per month over 2019–2021. The post-ChatGPT gap shows the opposite. It survives the excision of its own crisis window, and its factor-adjusted alpha is negative at -30 basis points per month, although that residual is imprecisely estimated. The pre-period gap therefore reflects the factor structure of high-markdown firms during a growth rally, whereas the post-period gap exceeds what the same structure implies.

Two hand-picked windows cannot say how unusual a gap of this size is for this design, so Appendix Table A.15 and Figure A.1 run the identical specification over every thirty-seven-month

window that closes on or before October 2022. Sample definition, exposure sorts, pre-event subsector fixed effects, winsorization rule, and frozen controls are held fixed across windows, and the post-ChatGPT window is estimated through the same code path so that the coefficients are exactly comparable. The 58 pre-ChatGPT windows deliver high-markdown coefficients with a median of -8.3 points, a standard deviation of 5.7 points, and a most negative value of -18.6 points. The post-ChatGPT estimate of -25.0 points lies below all 58. With the frozen controls included the pre-ChatGPT distribution runs from -9.5 to $+10.6$ points, and the post-ChatGPT estimate of -20.5 points again lies below every window. Windows that overlap by up to thirty-six months are strongly dependent, so the resulting tail frequency is an empirical placebo benchmark rather than a size-controlled p-value. Read that way, it says that the post-ChatGPT differential sits outside the range this design generates in quiet periods, which is the clearest evidence that the 2019–2021 window does not simply reproduce the post-event result. Exposure and the frozen controls are measured after most placebo windows close, which is inherent to holding the treatment variable fixed across windows and is the convention the two hand-picked placebos already use.

We nonetheless caveat these results with three limitations of our identifying assumptions. First, the estimates are cross-sectional associations between pre-event occupational structure and post-event outcomes, disciplined by placebos and fixed effects rather than by random assignment. Second, the setting is a single industry with roughly two dozen subsector clusters, which is why a wild-cluster bootstrap p-value accompanies every headline number and no long-horizon table carries stars. Third, all firms share one realized event window, so the rolling benchmark rather than the cross-sectional standard error is what disciplines the window-specific component of the uncertainty. We interpret the evidence as showing that institutional constraints carry return information beyond technical exposure. The design does not identify a causal effect of AI repricing.

5.5 Profitability and employment after ChatGPT

If the market is pricing deployable AI rather than sentiment, the repricing should appear in cash flows with a lag. Table 7 tests this in Compustat firm-year data with firm and year fixed effects, comparing fiscal years 2023 and later with the pre-2023 period across the same pre-event exposure sorts. A one-standard-deviation increase in deployable exposure is associated with a post-2022 increase of 0.5 percentage points in EBITDA over assets and 0.4 points in net income over assets, alongside a 4.2 percent decline in employees. Revenue per employee moves in the same direction without reaching significance. High-markdown firms show the opposite pattern in columns (5) through (8): the profitability interactions are zero to three decimal places, while employment falls by 10.3 percent per standard deviation of the markdown.

The Revelio staffing data, which extend through mid-2026, corroborate the employment margin. Firms with high deployable exposure show a flat pre-trend through 2022 and then lose four to six log points of relative employment through 2026, a trajectory break timed at the release, while the employment declines associated with technical exposure and the markdown continue pre-existing trends and therefore support no causal reading.

Firms whose AI-exposed tasks are deployable converted the technology into margins, partly by operating with fewer workers, and their valuations moved with the change. Firms whose exposure is marked down by governance constraints shed employment without a compensating profitability gain, and in the relative-wealth specification of Appendix Table A.14 they underperformed low-markdown firms by 20.0 percentage points. The margin along which returns and employment differentiate finance firms over this period is deployability rather than technical feasibility.

Table 7: Fundamentals after ChatGPT: profitability and employment

	EBITDA/A	NI/A	Rev/Emp	ln(Emp)	EBITDA/A	NI/A	Rev/Emp	ln(Emp)
Deployable (std.) × post-2023	0.005**	0.004**	91.246	-0.042*				
	(0.002)	(0.002)	(62.641)	(0.024)				
Markdown (std.) × post-2023					0.000	-0.000	70.004	-0.103***
					(0.002)	(0.002)	(62.344)	(0.020)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5,082	5,282	5,105	5,108	5,082	5,282	5,105	5,108
R-squared	0.78	0.69	0.66	0.97	0.78	0.68	0.66	0.97

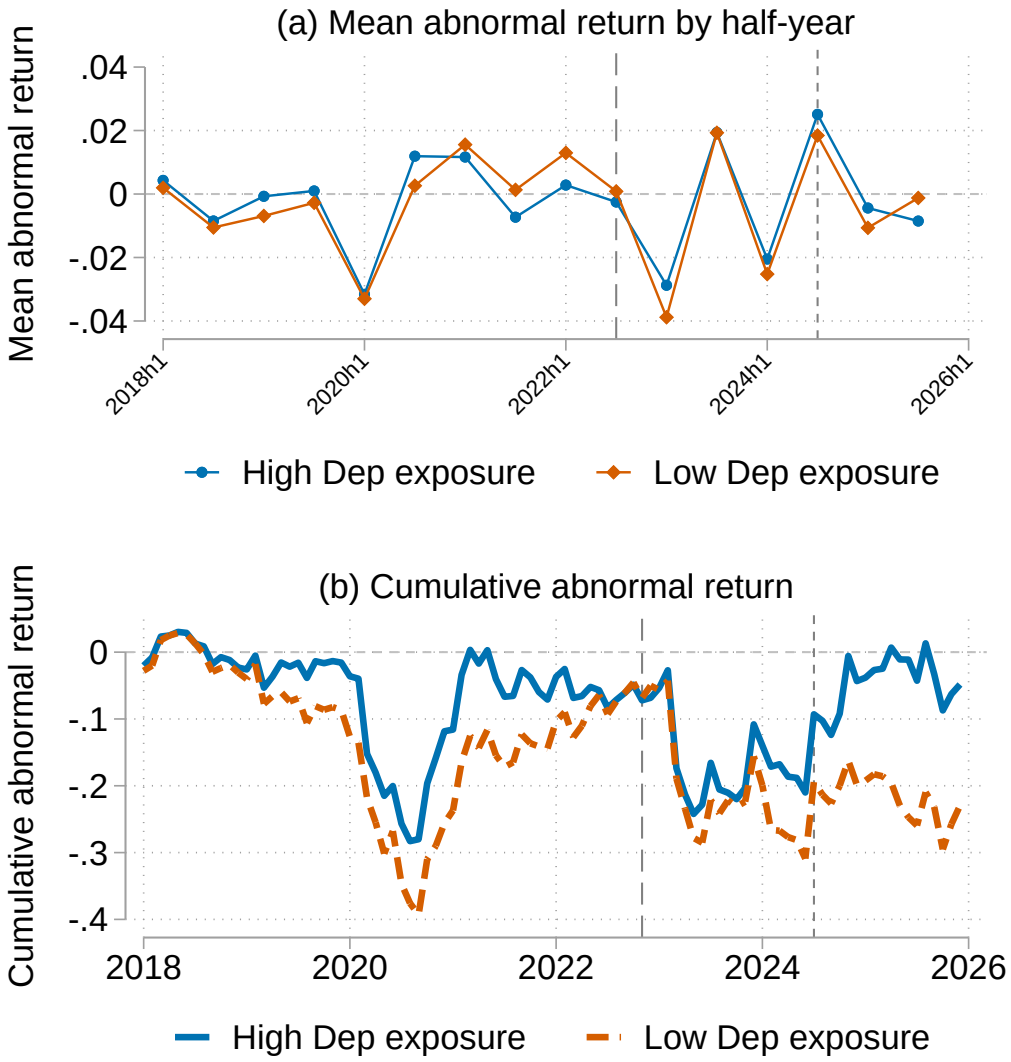
Notes. Compustat firm-year panel, fiscal years 2015–2025, for the return-sample firms. Each column regresses the outcome named in the header on a standardized pre-ChatGPT exposure measure interacted with an indicator for fiscal years 2023 and later, with firm and fiscal-year fixed effects. Columns (1)–(4) use deployable exposure and columns (5)–(8) the markdown. Ratio outcomes are winsorized at the 1st and 99th percentiles within year. Standard errors are clustered on firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6 Conclusion

This paper separates what frontier models can do to finance tasks from what financial institutions can deploy. Across 2,199 O*NET tasks, 99 finance-and-insurance occupations, and eight frontier model checkpoints, adding finance context and named governance regimes reduces measured exposure by 12.0 employment-weighted points on a 0–100 scale, about one-fifth of the mean feasibility score, and the markdown is positive for every model in the panel. It is largest in regulated, client-facing occupations such as loan interviewers, credit analysts, credit authorizers, underwriters, tellers, and loan officers, and smallest in support, marketing, software, and maintenance roles. The feasibility rung of the measure validates against existing GPT-based O*NET exposure scores, while the markdown correlates only 0.23 with those scores. Models also agree more about technical feasibility than about finance deployability, so the omitted institutional layer produces both a level adjustment and instability in the exposure construct.

The paper contributes a measurement object. Technical feasibility and deployable exposure answer different economic questions, and treating them as the same object overstates exposure

for tasks whose outputs must pass through review, explanation, confidentiality controls, records retention, validation, or accountable human decision-making. The overstatement is largest precisely where existing exposure measures rank finance work as most exposed. For labor-market measurement, this changes which occupations are predicted to adjust first. For asset pricing, it changes which firms carry AI exposure: firm-level exposure built from staffing shares and feasibility scores assigns the same exposure to firms whose task bundles differ sharply in deployability. Using Revelio Labs firm-by-occupation staffing shares, we test whether markets price the deployable component around AI capability shocks and in the cross-section of expected returns. The return evidence locates the pricing content in the institutional layer. Firms in the top half of the markdown distribution underperform the bottom half by roughly 25 percentage points in market-adjusted cumulative abnormal returns over the three years after ChatGPT, a differential that lies below every estimate the same design produces in the 58 pre-ChatGPT windows of equal length. It arrives in the first half of 2023 and persists rather than accumulating with later capability releases, and the pattern is corroborated in fundamentals, with high-markdown firms shedding employment without a profitability gain. With one realized event window and about two dozen subsector clusters, we read this as evidence that institutional constraints carry return information beyond technical exposure rather than as a causal estimate of AI repricing.



Panel (a): equal-weighted mean market-adjusted return, winsorized 1/99 within month. Panel (b): the unwinsorized monthly group means, summed over months — the group analogue of the firm-level CAR in the table. Dashed: ChatGPT release (2022m11). Short-dash: agentic tools (2024h2).

Figure 3: Abnormal returns for firms above and below the median of pre-ChatGPT deployable exposure Inst^F. Panel (a) plots the equal-weighted mean market-adjusted monthly return by calendar half-year, winsorized at the first and ninety-ninth percentiles within month. Panel (b) sums the unwinsorized monthly group means from January 2018. It is the group analogue of the firm-level cumulative abnormal return in Table 3, which is measured over the shorter post-ChatGPT window and sorts firms on the markdown rather than on deployable exposure. The dashed line marks the release of ChatGPT in November 2022 and the short-dashed line the arrival of agentic tools in the second half of 2024. The gap between the groups is not adjusted for subsector composition; see the text.

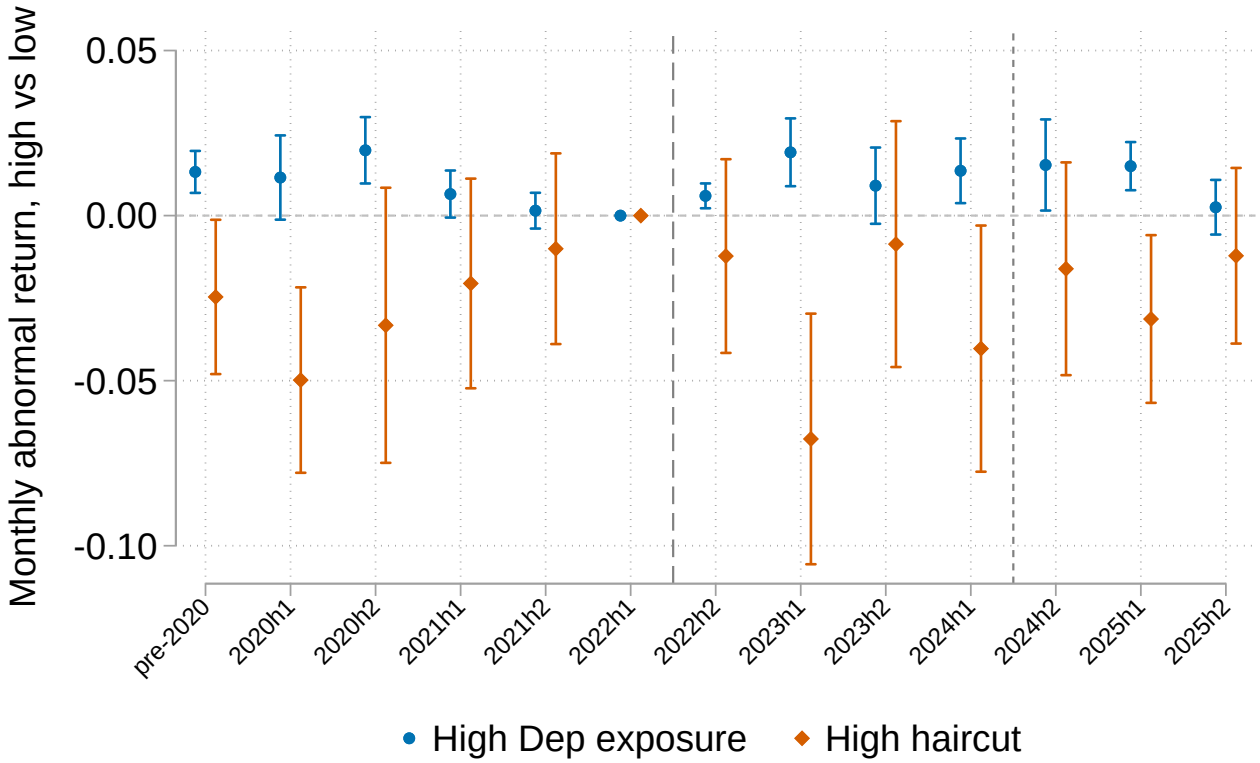


Figure 4: Event-study timing of the exposure differential. Each point is the coefficient on a high-exposure indicator interacted with a calendar half-year, from a regression of monthly abnormal returns with firm and month fixed effects, relative to the first half of 2022. Circles mark the deployable-exposure sort (Inst^F, labeled Dep) and diamonds the markdown sort ($\Delta^{F,AC}$, labeled haircut). Bars are 95% confidence intervals from standard errors clustered two-way on firm and month.

Tables and Figures

Table 8: Mean exposure by model and prompt rung

Model	Rung A	Rung B	Rung C	Haircut $A - C$
claude-sonnet-4	57.0	43.9	39.8	17.2
claude-sonnet-4.6	52.0	41.6	40.2	11.8
claude-opus-4.8	58.0	45.6	46.7	11.2
gpt-5.5	69.7	56.8	59.4	10.3
claude-sonnet-5	47.2	36.1	38.1	9.0
gpt-4o	51.1	49.4	46.2	4.9
gpt-3.5-turbo	65.7	56.3	61.3	4.4
gpt-4.1	60.5	54.4	59.3	1.2

Notes. Each cell is the mean exposure score, rescaled to a 0–100 range, averaged across the 2,199 finance-and-insurance task statements for the given model. Rung A is bare technical feasibility, rung B adds finance-industry framing, and rung C adds named finance governance regimes. The final column is the within-model institutional markdown $A - C$. Every model exhibits a positive markdown, from 1.2 to 17.2 points. The panel covers rungs A–C.

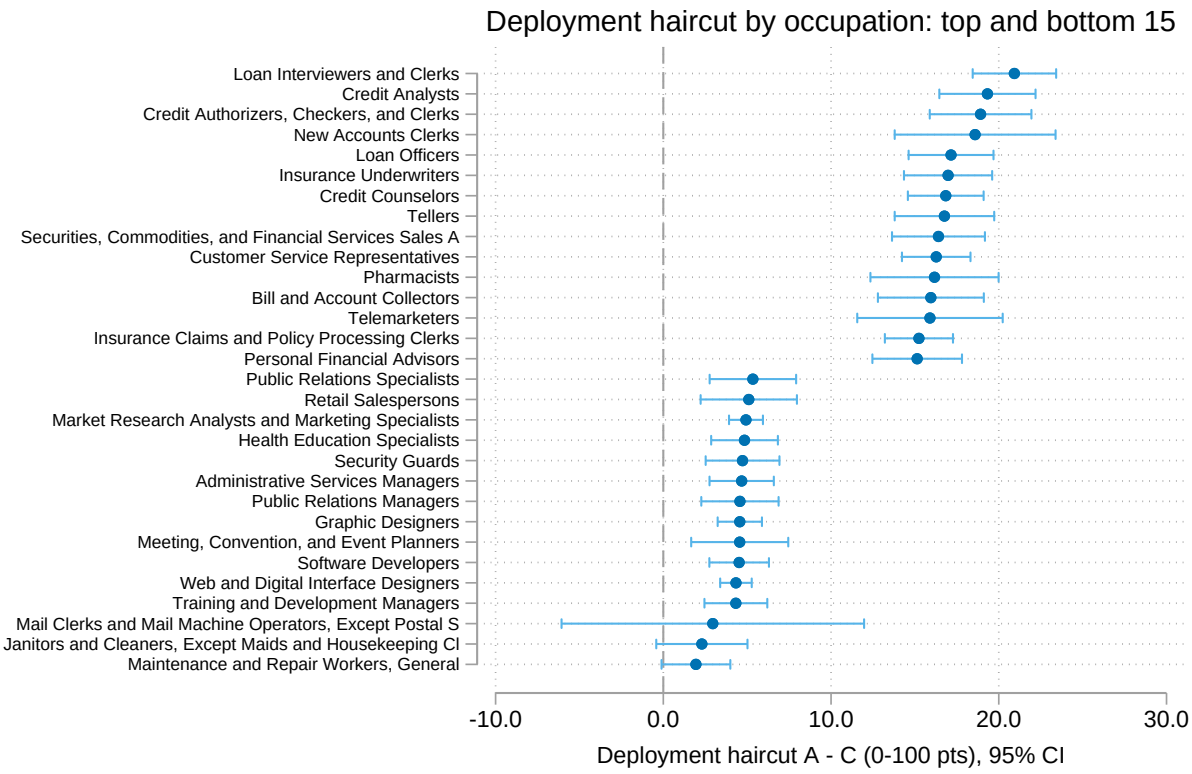


Figure 5: Institutional markdown by occupation, with 95% confidence intervals. Each row is one of the fifteen most- and fifteen least-affected six-digit SOC occupations, ordered by the pooled institutional markdown $A - C$ (0–100 points). Points are importance-weighted markdowns averaged over the eight models, and horizontal caps are 95% confidence intervals from standard errors clustered on task. Regulated, client-facing occupations sit at the top. Support, software, marketing, and maintenance roles sit at the bottom.

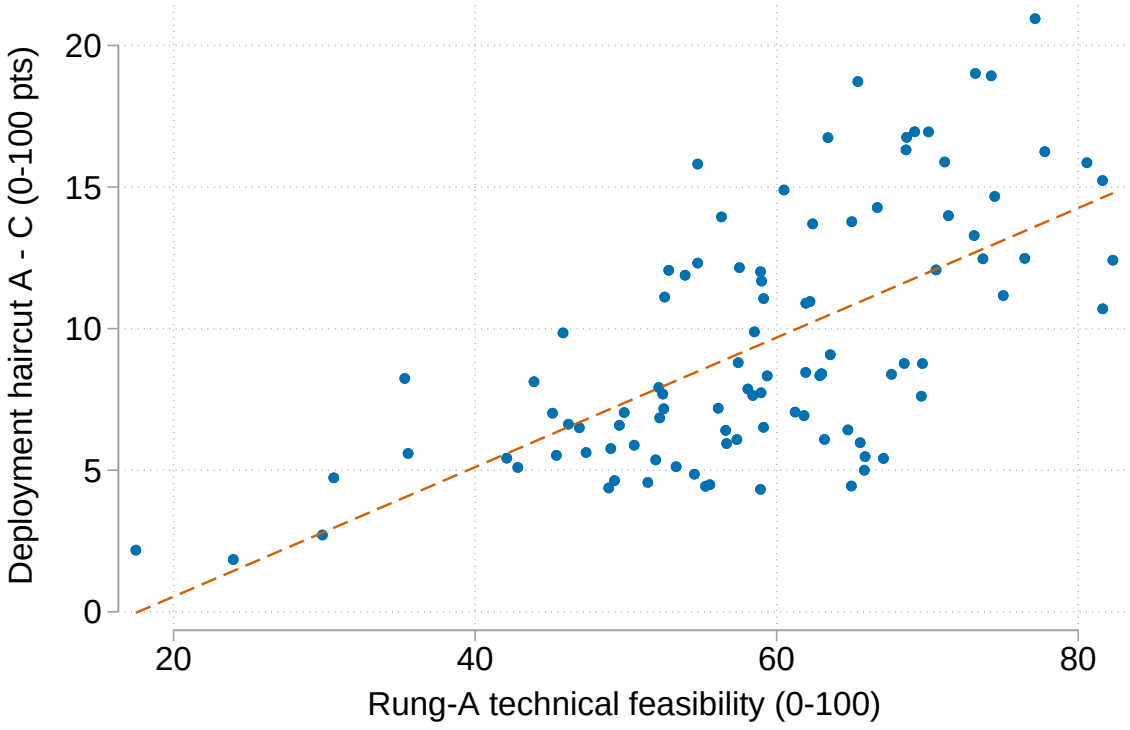


Figure 6: The institutional markdown against technical feasibility across occupations. Each point is one of the 99 finance-and-insurance occupations, plotting the mean institutional markdown $A - C$ (vertical axis, 0–100 points) against mean rung-A technical feasibility (horizontal axis, 0–100 points). The line is an OLS fit. The positive slope identifies the "capable but not deployable" region: occupations that are technically exposed also tend to lose more exposure once finance constraints are introduced.

Table 9: Institutional markdown by model, importance-weighted

Model	A	B	C	$A - C$	CI low	CI high
claude-sonnet-4	57.7	45.3	39.1	18.6	16.6	20.5
claude-sonnet-4.6	53.3	42.5	40.3	13.0	11.9	14.1
claude-opus-4.8	58.6	46.2	46.7	11.9	11.0	12.9
claude-sonnet-5	48.8	37.0	37.9	11.0	9.7	12.2
gpt-5.5	71.5	58.2	60.7	10.8	10.0	11.5
gpt-4o	53.1	51.0	47.8	5.3	4.7	6.0
gpt-3.5-turbo	66.5	60.0	62.4	4.1	2.7	5.5
gpt-4.1	62.2	56.3	60.9	1.3	0.5	2.0

Notes. For each model, columns A–C report mean exposure (0–100) weighted by O*NET task importance within occupation. Column $A - C$ is the institutional markdown with a 95% confidence interval from standard errors clustered on the 99 SOC occupations. The markdown is positive and its confidence interval excludes zero for all 8 of 8 models. Pooled over models and weighted by NAICS-52 sector employment, the economy-wide markdown is 12.0 points (95% CI [9.6, 14.8]). The importance-weighted, equal-occupation markdown is 9.5 points. The mean $A - B$ adjustment is 10.1 points, compared with 1.9 points for $B - C$.

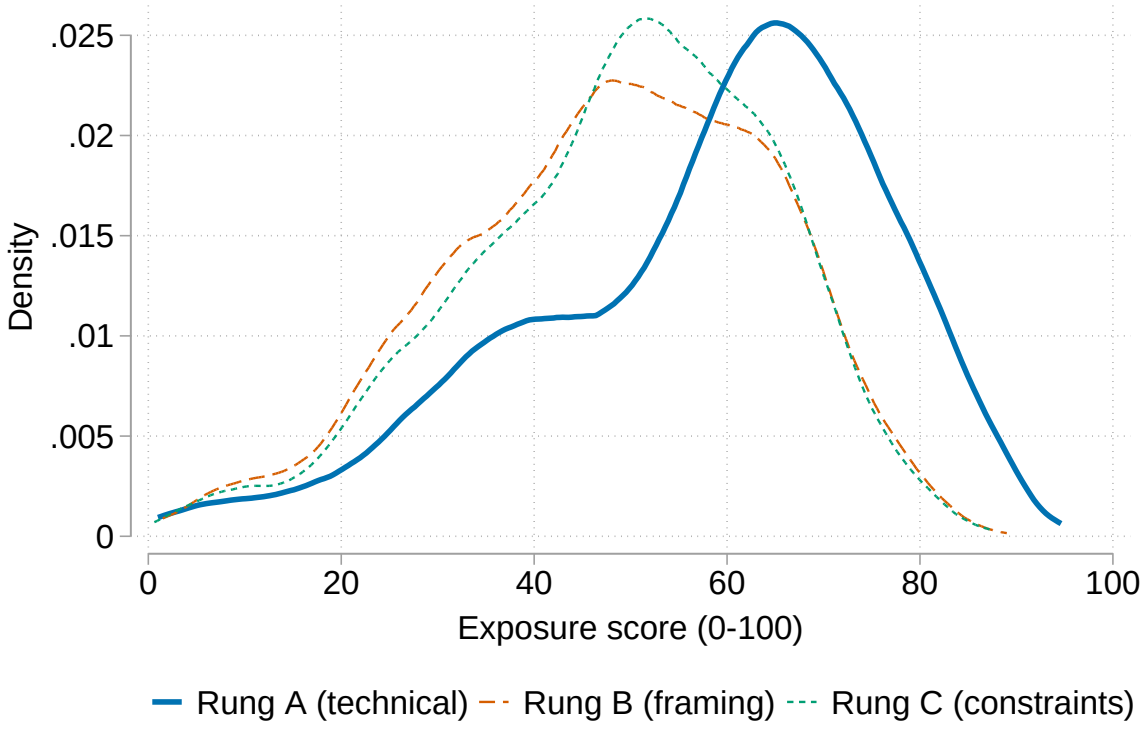


Figure 7: Task-level distribution of exposure scores by prompt rung. Kernel densities of task-mean exposure scores (0–100), pooled across models, at rung A (technical feasibility), rung B (finance framing), and rung C (named governance regimes). Adding finance context shifts the distribution left, which is the distributional signature of the institutional markdown.

Table 10: Institutional markdown by occupation: fifteen most- and fifteen least-affected occupations

Occupation (SOC6)	Tech (A)	$A - C$	CI low	CI high
Loan Interviewers and Clerks	77.2	20.9	18.4	23.4
Credit Analysts	75.3	19.3	16.5	22.2
Credit Authorizers, Checkers, and Clerks	73.3	18.9	15.9	21.9
New Accounts Clerks	65.6	18.6	13.8	23.4
Loan Officers	69.2	17.2	14.6	19.7
Insurance Underwriters	70.2	17.0	14.3	19.6
Credit Counselors	68.9	16.8	14.6	19.1
Tellers	63.4	16.8	13.8	19.7
Securities, Commodities, and Financial Services Sales A	68.4	16.4	13.6	19.2
Customer Service Representatives	77.6	16.3	14.2	18.3
Pharmacists	55.5	16.2	12.3	20.0
Bill and Account Collectors	71.1	16.0	12.8	19.1
Telemarketers	80.6	15.9	11.6	20.2
Insurance Claims and Policy Processing Clerks	81.5	15.2	13.2	17.3
Personal Financial Advisors	60.8	15.1	12.5	17.8
Public Relations Specialists	53.5	5.3	2.8	7.9
Retail Salespersons	43.1	5.1	2.2	8.0
Market Research Analysts and Marketing Specialists	66.3	4.9	3.9	5.9
Health Education Specialists	54.5	4.8	2.9	6.8
Security Guards	30.8	4.7	2.5	6.9
Administrative Services Managers	49.3	4.7	2.8	6.6
Public Relations Managers	51.5	4.6	2.3	6.9
Graphic Designers	65.0	4.6	3.2	5.9
Meeting, Convention, and Event Planners	49.0	4.6	1.7	7.4
Software Developers	55.8	4.5	2.7	6.3
Web and Digital Interface Designers	58.8	4.3	3.4	5.3
Training and Development Managers	55.2	4.3	2.5	6.2
Mail Clerks and Mail Machine Operators, Except Postal S	29.8	3.0	-6.1	12.0
Janitors and Cleaners, Except Maids and Housekeeping Cl	17.7	2.3	-0.4	5.0
Maintenance and Repair Workers, General	24.4	1.9	-0.1	4.0

Notes. Each row is a six-digit SOC occupation. "Tech (A)" is the importance-weighted rung-A technical-feasibility score (0–100), and "$A-C$" is the occupation's institutional markdown, pooled over the eight models and importance-weighted over tasks. Confidence intervals use standard errors clustered on task. Occupations are sorted on the markdown. Regulated, client-facing roles carry the largest markdowns, while support, marketing, software, and maintenance roles carry the smallest. Figure 5 plots the same comparison with confidence intervals.

Table 11: Cross-model agreement by prompt rung

Rung	Mean pairwise Spearman
A	0.804
B	0.768
C	0.738

Notes. Mean pairwise Spearman rank correlation of task scores across the eight models, computed within each rung. Agreement is highest at rung A and declines as finance context is added. The models therefore agree more about technical feasibility than about deployability inside finance.

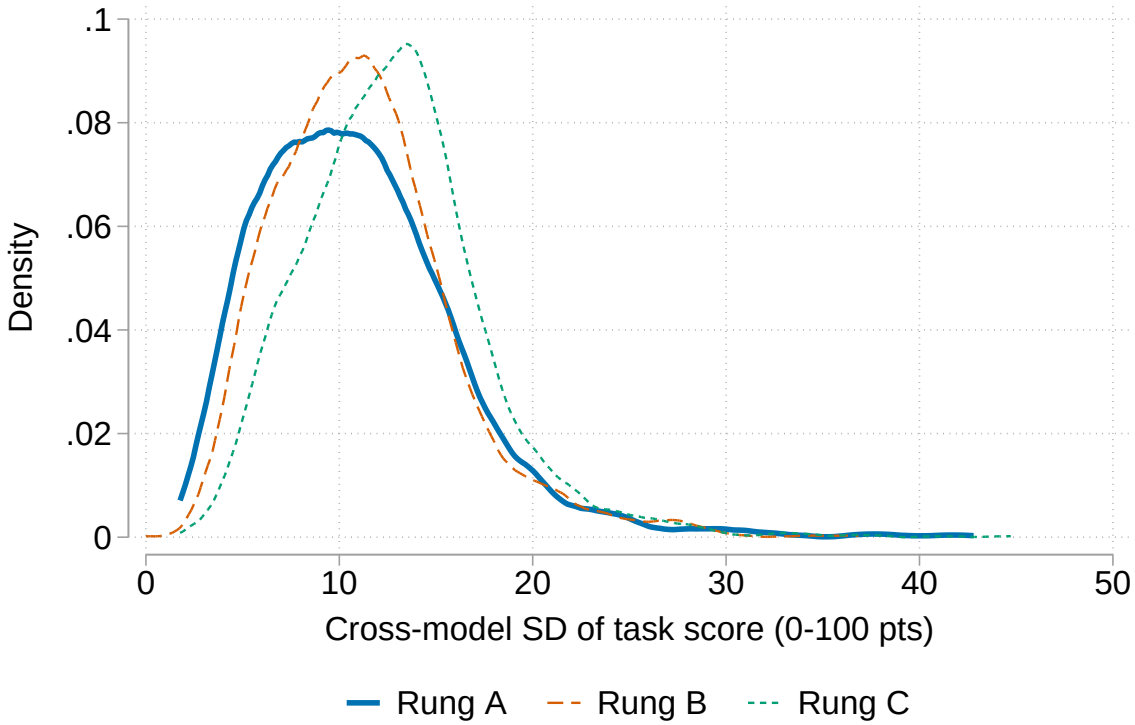


Figure 8: Cross-model dispersion by prompt rung. Kernel densities of each task's cross-model standard deviation (0–100 points) at rungs A, B, and C. Disagreement widens as institutional context is added, consistent with the view that models agree more on bare capability than on deployability inside finance.

A Coding Guide for Finance-Industry Institutional Constraints

This appendix defines the five dimensions of the co-dependence vector H_k^F introduced in Section 3. The unit of coding is an O*NET task performed inside a finance-industry firm. The current paper reports rungs A–C, so the codebook anchors the interpretation of the construct rather than the generation of the main tables. It also disciplines the interpretation of the model-generated diagnostics reported below.

The codes measure deployment burden under the baseline human-in-the-loop finance deployment mode. They do not measure whether AI is legally prohibited. They also do not measure whether a firm would adopt AI for the task. A code of 2 means that compliant deployment would require substantial institutional scaffolding. It does not mean deployment is impossible.

A.1 General coding rules

Coders apply the following rules in order.

1. **Code the task in finance, not the occupation title.** A generic task can receive a high code if performing it inside a financial institution makes customer data, supervised communication, model governance, or accountability central to deployment.

2. **Code marginal deployment burden.** Ask how much additional governance, review, retention, explanation, or accountability is required before AI output can be used in production.

3. **Separate drafting from action.** A task that drafts an internal note is usually lower coded than a task that recommends, triggers, or explains a customer, trading, credit, insurance, regulatory, or financial-reporting action.

4. **Use 2 only when the constraint is intrinsic to the task.** A task receives 2 when the constraint is not incidental but part of the task's production function inside finance.

5. **Resolve ambiguity conservatively.** If a task could be performed with public or nonpublic data, assume the ordinary finance-industry version. If the task statement is too vague to determine whether a constraint binds, assign 1 rather than 2 unless the occupation and task family clearly imply the stronger code.

A.2 Model-risk governance: $h_k^{MRM,F}$

This dimension captures whether AI use for the task would plausibly fall inside model-risk governance or a comparable internal model-control process. The anchor is not a claim that every LLM use is formally in scope of banking model-risk guidance. The anchor is the governance burden associated with recurring analytic systems that feed material finance decisions: inventory, testing, validation, monitoring, change control, vendor oversight, and clear ownership.

Code 0: not binding. The AI output is one-off, clerical, editorial, or informational and does not feed a recurring quantitative decision process. Examples include summarizing a public filing for internal reading, drafting a generic meeting agenda, rewriting an internal training paragraph, or extracting public facts for an analyst who independently verifies the result.

Code 1: partially binding. The task uses AI as recurring analytic assistance or produces an intermediate input to a finance decision, but a human expert reviews the output and the task is not itself a governed model or decision engine. Examples include drafting risk commentary from analyst notes, summarizing stress-test narratives prepared elsewhere, classifying customer complaints for human review, or preparing a first-pass exception report that an employee validates.

Code 2: strongly binding. The task produces, updates, validates, monitors, or operationalizes a model or analytic system used in credit, pricing, capital, valuation, market risk, liquidity risk, fraud, AML, insurance underwriting, claims reserving, financial reporting, or other material finance decisions. Examples include developing credit-score model features, monitoring model drift, generating risk-limit recommendations, triaging AML alerts at scale, or producing automated valuation inputs used in books and records.

Boundary cases. Writing documentation for an already governed model is usually 1, not 2, unless the AI system also generates or changes model outputs. Simple arithmetic or deterministic rule application is usually 0 or 1 unless it becomes part of a recurring decision system. Vendor-provided AI tools are coded by the task they perform, not by whether the vendor bears some governance responsibility.

A.3 Adverse-action and explainability: $h_k^{AA,F}$

This dimension captures whether the task is tied to credit approval, denial, pricing, limit setting, account termination, or other actions for which the financial institution must be able to explain the decision to an applicant or customer. The core anchor is ECOA and Regulation B for credit

decisions. Insurance and investment-advice tasks may require explanations or suitability documentation, but they are coded under this dimension only when the task is credit-like or explicitly tied to adverse-action reasoning. Otherwise they are captured by the supervision, confidentiality, or accountability dimensions.

Code 0: not binding. The task is unrelated to credit eligibility, credit terms, account reduction, or customer-specific denial. Examples include summarizing a 10-K, preparing a macroeconomic briefing, reconciling internal ledgers, drafting generic marketing copy, or analyzing public market data.

Code 1: partially binding. The task supports a credit or credit-adjacent decision but does not itself recommend, determine, or explain the action. Examples include collecting loan-application information, summarizing borrower financials for an underwriter, flagging missing documentation, preparing preliminary risk factors, or drafting an internal credit memo section that a human rewrites before decision.

Code 2: strongly binding. The task recommends, determines, justifies, documents, or communicates approval, denial, pricing, credit-line reduction, counteroffer terms, or other credit actions where reason codes, explanations, or reviewable decision files are central. Examples include drafting an adverse-action notice, selecting denial reasons, recommending loan approval or denial, assigning risk-based pricing reasons, or explaining a credit-line decrease.

Boundary cases. A general fraud-screening task is usually 1 if it may affect account access and 2 if it directly supports closure, denial, or account restriction. Commercial-credit tasks are

coded by whether the task requires explainable decision documentation, even if the formal notice regime differs from consumer credit. Pure affordability education or financial coaching is 0 unless it becomes part of a credit decision.

A.4 Confidentiality and data restrictions: $h_k^{CONF,F}$

This dimension captures whether the task requires customer, account, transaction, credit-file, insurance, health, employee, trading, or proprietary business information whose use changes the feasible AI deployment architecture. The anchor is the finance industry's obligation to safeguard customer information and control sensitive internal data.

Code 0: not binding. The task can be performed using public, synthetic, or non-sensitive information. Examples include summarizing public company filings, drafting generic policy language, reviewing public interest-rate data, preparing public-market commentary for internal use, or extracting facts from publicly available regulatory releases.

Code 1: partially binding. The task uses internal or potentially sensitive information, but not detailed customer-level nonpublic financial information. Examples include summarizing internal policies, analyzing aggregated portfolio statistics, preparing management reports from de-identified data, drafting internal process documentation, or reviewing employee-facing procedures.

Code 2: strongly binding. The task requires nonpublic customer or account-level information, transaction histories, credit reports, loan files, claim files, personally identifiable information, health-related insurance information, authentication credentials, suspicious-activity data, or other sensitive records. Examples include reviewing a mortgage application, analyzing transaction pat-

terns for fraud, summarizing a customer complaint file, preparing an insurance-claim recommendation, or generating customer-specific investment advice from account holdings.

Boundary cases. Public filings remain 0 even if the analysis is valuable. Proprietary research without customer data is usually 1. Aggregated data are 1 unless the task requires re-identification risk, small-cell information, or drill-down to customer records, in which case the code is 2. If the task can be completed by sending sensitive data to an external model only after redaction or secure-environment controls, code the unredacted ordinary task and record the control in coder notes.

A.5 Supervisory review and recordkeeping: $h_k^{SUP,F}$

This dimension captures whether the task produces communications, recommendations, marketing material, research, or other records that must be supervised, approved, retained, or reconstructed. The anchors include broker-dealer communications rules, investment-adviser marketing and books-and-records rules, and analogous internal retention regimes.

Code 0: not binding. The task produces internal work product with no client, investor, regulator, or public distribution and no special retention requirement beyond ordinary business records. Examples include an internal summary of a public article, a draft data dictionary, a private analyst note not distributed externally, or a temporary internal checklist.

Code 1: partially binding. The task produces material that may support external communication or must be retained, but it is not itself a retail communication, recommendation, advertisement, research distribution, or final client communication. Examples include drafting a

client-meeting outline for adviser review, preparing institutional background material, summarizing a research report for internal distribution, or classifying communications for later supervisory review.

Code 2: strongly binding. The task produces or materially shapes a customer-, investor-, regulator-, or public-facing communication that requires review, approval, filing, retention, or reconstruction. Examples include drafting retail investment recommendations, preparing marketing material with performance claims, generating broker-dealer communications with the public, writing customer complaint responses, preparing research for distribution, or producing regulated adviser advertisements.

Boundary cases. Internal communications are not automatically 0. Code 1 if they are part of a review chain for external material. Purely factual service notices are usually 1 unless they include advice, recommendations, performance, pricing, denial, or product promotion. Social-media, website, and mass-email tasks are usually 2 when they promote financial products or services.

A.6 Auditability and human accountability: $h_k^{ACC,F}$

This dimension captures whether the task requires an accountable human, a defensible audit trail, or escalation to a licensed, supervisory, fiduciary, or control-function employee before the output can be acted upon. It is broader than formal regulation. It measures whether finance production requires the institution to know who reviewed, approved, changed, or relied on the AI output.

Code 0: not binding. The task is low-stakes, reversible, and does not require a named accountable decision-maker. Examples include formatting internal notes, drafting a generic glossary, summarizing public background material, or organizing non-sensitive files.

Code 1: partially binding. Human review is expected and a review trail is useful, but the AI output is not itself a final or near-final action. Examples include preparing a preliminary audit workpaper, drafting an internal risk summary, classifying support tickets for routing, or generating a checklist for a compliance analyst.

Code 2: strongly binding. The task produces, recommends, approves, or documents an action for which a specific human, committee, licensed professional, fiduciary, supervisor, or control owner remains responsible. Examples include approving a loan, recommending a trade, issuing investment advice, signing off on financial statements, closing a suspicious-activity investigation, settling an insurance claim, or approving a client-facing recommendation.

Boundary cases. Drafting is usually 1 unless the draft is the operative record or the task statement says the worker approves, recommends, certifies, signs, or decides. Escalation tasks are 2 when failure to escalate creates material customer, market, regulatory, or financial-reporting risk. Tasks performed by licensed personnel are not automatically 2. The task must require the licensed person's accountability.

A.7 Example task codes

Table A.1 gives illustrative codes. The table is not a substitute for the rules above. It is intended to discipline borderline cases before coders begin.

Table A.1: Illustrative finance-industry constraint codes

Task performed inside a finance firm	MRM	AA	CONF	SUP	ACC	Coding rationale
Summarize a public 10-K for internal analyst review	0	0	0	0	0	Public input, internal use, no customer action.
Draft an adverse-action explanation for a denied consumer loan	1	2	2	1	2	Credit decision file, customer data, reason-giving, accountable approval.
Prepare a retail investment recommendation email	0	0	2	2	2	Customer holdings and risk profile, supervised communication, accountable advice.
Classify suspicious transaction alerts for AML review	2	0	2	0	2	Recurring risk system, sensitive transaction data, escalation accountability.
Reconcile internal ledger entries	1	0	1	0	1	Internal financial controls. Review needed but no external communication.
Generate insurance-claim settlement recommendation	1	0	2	1	2	Claim file data, customer communication, accountable settlement action.
Draft generic cybersecurity training for employees	0	0	1	0	0	Internal policy context but no customer decision or regulated output.
Prepare marketing copy showing fund performance	0	0	0	2	2	Performance advertising and approval/retention requirements.

A.8 Human calibration materials

The replication package retains the two author-coded files from the original 360-task pilot as an internal benchmark. Those files predate the expansion to the current 2,199-task universe and are not used to construct the main exposure measures or as a validation sample for the current universe. The main results therefore do not rely on an independently coded H_k^F vector.

B Scoring Configuration and AI-Workflow Disclosure

This appendix records the scoring configuration and the AI-workflow documentation required by the venue for which the project is prepared. The exposure panel is scored over 2,199 O*NET task statements across 99 finance-and-insurance occupations, by eight frontier model checkpoints from two independent developers–OpenAI (GPT) and Anthropic (Claude)–accessed through a common API router. The scored rungs are A–C: bare technical feasibility, finance-industry framing, and named finance governance regimes.

Each task–model–rung cell is scored five times at temperature 0.7 with deterministic per-repetition seeds and summarized by the median. The score is elicited as an integer from 0 to 100 against the rubric reproduced in the replication package, under a structured-output contract that degrades gracefully from strict JSON-schema enforcement to prompt-embedded schemas for models without strict support. Every call logs the resolved model identifier, serving provider, seed, token usage, cost, and a hash of the exact prompt. A campaign manifest pins the rubric version, prompt-template hash, and checksum of the relevant input files, so that drift in prompts, codebook, or model versions is detectable ex post.

The paper itself is produced under a documented human–AI workflow. The replication package contains the verbatim transcripts of AI conversations used in the project, including initial prompts, human review and editing steps, and a mechanical accounting of the fraction of project lines contributed by the authors versus AI assistants. The measurement design, interpretation of results, constraint-codebook rules, and final manuscript judgments are the authors' own.

C Additional Tables and Figures

This appendix contains exhibits that are useful for audit and robustness but are not main results. The model-generated co-dependence proxy is especially important to quarantine: it is built from the rung-C model responses and therefore cannot be treated as an independent validation of the rung-C score.

Table A.2: Distribution of the primary binding constraint at rung C

Primary binding constraint (rung C)	Calls	Share (%)
ACC	29411	38.6
none reported	18029	23.6
CONF	16312	21.4
SUP	9392	12.3
MRM	2711	3.6
AA	424	0.6

Notes. Among rung-C model calls that completed successfully, the model names the single institutional constraint that most limits deployment. Accountability, confidentiality, and supervisory review dominate. Adverse-action and model-risk-management constraints are named rarely. About a quarter of calls report no binding constraint. This is a model self-diagnostic, not the independent adjudicated co-dependence code H_k^F.

Table A.3: The institutional markdown is not a verbosity or task-complexity artifact

	(1) Haircut $A - C$	(2) Tech (A)
Task word count	-0.0439*	-0.188***
	(0.0221)	(0.0612)
Wild-cluster p (word count)	0.064	0.004
Observations	17,592	17,592
Within R^2	0.174	0.150

Notes. Column (1) regresses the institutional markdown $A - C$ (0–100) on task word count and O*NET job-zone dummies with model fixed effects. Column (2) repeats the exercise for the rung-A technical score. Standard errors are clustered on SOC, and the wild-cluster bootstrap p-value is for the word-count coefficient. The word-count coefficient in the markdown regression is -0.044 points per word (wild-cluster $p = 0.064$), so longer task statements do not mechanically receive larger markdowns. Longer tasks also receive slightly lower technical scores. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Robustness of the pooled institutional markdown

Specification	Haircut	CI low	CI high
Signed A-C (baseline)	12.02	10.54	13.50
Nonnegative max(A-C,0)	12.53	11.15	13.91
Proportional (A-C)/A	17.95	16.11	19.79
Leave-out claude	7.17	6.22	8.12
Leave-out gpt	16.87	14.66	19.08
Drop adherence outliers	12.31	10.80	13.82

Notes. Each row re-estimates the economy-wide, employment-weighted pooled markdown under an alternative definition or sample, with standard errors clustered on SOC and a 95% confidence interval. The signed baseline is $A - C$. The nonnegative variant censors negative task-level markdowns at zero. The proportional variant is $(A - C)/A$. The leave-out rows drop one developer family at a time, and the last row drops the two rung-C instrument-adherence outliers flagged in the audit.

Table A.5: Construct-validity diagnostic using the model-generated co-dependence proxy

	(1) Δ (A−C)	(2) Δ (A−C)	(3) Tech (rung A)
H (LLM proxy)	10.94***		23.89***
	(0.794)		(3.763)
MRM share		15.84***	
		(1.668)	
AA share		21.16***	
		(4.193)	
CONF share		11.61***	
		(1.135)	
SUP share		9.918***	
		(0.999)	
ACC share		10.25***	
		(0.762)	
Wild-cluster p (H)	0.000		0.000
Observations	15,256	15,256	15,256
Within R^2	0.202	0.205	0.214

Notes. Column (1) regresses the institutional markdown $A - C$ (0–100) on the model-generated co-dependence proxy H_{llm}. Column (2) replaces H_{llm} with its five diagnostic shares. Column (3) is the placebo regression of rung-A technical exposure on H_{llm}. These estimates are pipeline diagnostics, not main results. The proxy is built from the same rung-C responses that define the constrained score, so the estimates carry a mechanical-circularity component. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: Cross-model disagreement and the model-generated co-dependence proxy

	(1) Disp_A	(2) Disp_A−Disp_C
H (LLM proxy)	-6.236***	-3.569***
	(1.224)	(0.572)
Wild-cluster p (H)	0.000	0.000
Observations	1,907	1,907
Within R^2	0.063	0.063

Notes. Column (1) regresses the cross-model standard deviation of the rung-A task score on the model-generated co-dependence proxy. Column (2) uses the $A - C$ compression of dispersion. Both include task controls and occupation fixed effects, with standard errors clustered on SOC and wild-cluster bootstrap p-values. The negative coefficient is a warning sign for the proxy: under this diagnostic measure, high-H tasks are not the tasks with the highest bare-prompt dispersion. This table should be re-estimated with the independent H_k^F vector before being interpreted substantively. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Robustness of the model-generated co-dependence gradient

	(1) Baseline	(2) Drop outliers	(3) Nonneg. Δ	(4) Proportional Δ	(5) $H\times$ family
H (LLM proxy)	10.94***	10.61***	8.671***	14.54***	14.54***
	(0.794)	(0.691)	(0.677)	(1.782)	(1.002)
claude					0
					(.)
gpt					0
					(.)
claude × Share of rung-C calls naming any binding constraint (LLM-diagnostic H proxy)					0
					(.)
gpt × Share of rung-C calls naming any binding constraint (LLM-diagnostic H proxy)					-7.208***
					(1.081)
Observations	15,256	11,442	15,256	15,046	15,256
Within R^2	0.040	0.040	0.034	0.009	0.047

Notes. Each column re-estimates the markdown-on-H_{llm} gradient. Column (1) is the two-way fixed-effects baseline, column (2) drops rung-C adherence outliers, columns (3)–(4) use nonnegative and proportional markdown definitions, and column (5) interacts the proxy with developer family. The same proxy caveat applies: these are pipeline checks, not paper results. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Exposure and the institutional markdown across model checkpoints

	(1) Tech (A)	(2) B	(3) C	(4) Δ (A−C)	(5) Δ, fam. trends	(6) Δ, drop outliers
Release date (years)	1.550	0.395	0.900	0.651	-6.984**	1.196
	(2.712)	(1.145)	(2.118)	(1.380)	(2.298)	(2.253)
GPT × release date					8.367**	
					(2.560)	
Wild-cluster p (model, Webb)	0.761	0.793	0.708	0.863	0.276	0.842
Observations	17,592	17,592	17,592	17,592	17,592	13,194
Within R^2	0.146	0.299	0.366	0.124	0.166	0.134

Notes. Columns (1)–(3) regress rung levels on checkpoint release date, measured in years, with task fixed effects and a developer-family dummy. Columns (4)–(6) use the institutional markdown $A - C$. The vintage regressor varies across only eight checkpoints in two families, so inference is a wild-cluster bootstrap at the model level with Webb weights. No rung level or pooled markdown trends significantly with release date. Because the rubric's "current frontier" anchor is indexical, level trajectories should be read as description rather than as a capability trajectory. Figure A.4 is the figure counterpart. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Usage behavior on high-markdown tasks

	(1) Augment. share	(2) Validation share	(3) AI autonomy
Markdown $\Delta^{F,AC}$	1.747**	0.443***	-0.0955***
	(0.771)	(0.144)	(0.0141)
Technical exposure Tech^F	-3.221***	-0.393*	0.0378**
	(0.615)	(0.218)	(0.0181)
Wild-cluster p (markdown)	0.044	0.007	0.000
Observations	2,819	2,837	2,839
R^2	0.589	0.039	0.337

Notes. Complements Section 4.5. Each column regresses an Anthropic Economic Index conversation diagnostic on the markdown and technical exposure for published finance-universe tasks, pooling both usage sources and both months with source-by-month fixed effects. The outcomes are the share of the task's conversations in the augmentation collaboration bucket, the share showing the validation collaboration pattern, both in percentage points, and the mean assessed AI autonomy on a one-to-five scale. Exposure regressors are cross-model task means in standard-deviation units. Standard errors are clustered on occupation. The wild-cluster row reports the bootstrap p-value for the markdown coefficient. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10: Chat-versus-API wedges and the markdown

	(1) ln wedge	(2) ln wedge	(3) Auto. wedge
Markdown $\Delta^{F,AC}$	-0.0122	0.0341	3.469***
	(0.0598)	(0.0654)	(1.087)
Technical exposure Tech^F		-0.179***	-2.787***
		(0.0644)	(0.989)
Wild-cluster p (markdown)		0.656	
Observations	1,104	1,104	1,095
R^2	0.000	0.013	0.028

Notes. Complements Section 4.5. The sample is finance-universe tasks published in both Anthropic Economic Index sources in the same month, with month fixed effects. In columns (1)–(2) the dependent variable is the log of the task's chat usage share minus the log of its first-party API usage share, so a positive markdown coefficient would indicate that constrained tasks tilt toward individual adoption and away from production deployment. In column (3) the dependent variable is the task's automation collaboration share in API traffic minus the same share in chat traffic, in percentage points. Exposure regressors are cross-model task means in standard-deviation units. Standard errors are clustered on occupation. The wild-cluster row reports the bootstrap p-value for the markdown coefficient. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11: Occupation-level AI usage and exposure

	(1) ln usage	(2) ln usage	(3) ln usage
Technical exposure Tech^F	0.257		0.988***
	(0.178)		(0.192)
ln employment (OEWS NAICS-52)	0.118	0.127	0.309***
	(0.115)	(0.110)	(0.115)
Deployable exposure Inst^F		0.564***	
		(0.158)	
Markdown $\Delta^{F,AC}$			-1.177***
			(0.197)
Wild-cluster p (markdown)			0.000
Observations	386	386	386
R^2	0.033	0.108	0.289

Notes. Regression counterpart of Figure 2. The dependent variable is the log of the occupation's share of global Claude usage in a source and month, aggregated over the occupation's O*NET-SOC nodes. The sample is the published finance-universe occupations over two sources and two months. Occupation exposure at each rung is the importance-weighted mean over the occupation's tasks, averaged over the eight model checkpoints and standardized. All columns include source-by-month fixed effects and the log of NAICS-52 occupation employment from the OEWS. Standard errors are clustered on occupation. The wild-cluster row reports the bootstrap p-value for the markdown coefficient. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12: Sub-period timing of the exposure differential

	(1)	(2)	(3)	(4)	(5)	(6)
High Dep × 2022m12–2023m12	0.004					
	(0.003)					
High Dep × 2024m1–2024m12	0.005					
	(0.004)					
High Dep × 2025m1–2025m12	-0.001					
	(0.003)					
High Tech × 2022m12–2023m12		-0.007				
		(0.012)				
High Tech × 2024m1–2024m12		-0.004				
		(0.009)				
High Tech × 2025m1–2025m12		0.001				
		(0.005)				
High haircut × 2022m12–2023m12			-0.014			
			(0.013)			
High haircut × 2024m1–2024m12			-0.007			
			(0.011)			
High haircut × 2025m1–2025m12			-0.000			
			(0.007)			
Dep (std.) × 2022m12–2023m12				0.004*		
				(0.002)		
Dep (std.) × 2024m1–2024m12				0.004		
				(0.002)		
Dep (std.) × 2025m1–2025m12				0.001		
				(0.002)		
Tech (std.) × 2022m12–2023m12					-0.003	
					(0.005)	
Tech (std.) × 2024m1–2024m12					-0.000	
					(0.004)	
Tech (std.) × 2025m1–2025m12					0.000	
					(0.002)	
Haircut (std.) × 2022m12–2023m12						-0.007
						(0.007)
Haircut (std.) × 2024m1–2024m12						-0.004
						(0.006)
Haircut (std.) × 2025m1–2025m12						-0.001
						(0.004)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm + month	Firm + month	Firm + month	Firm + month	Firm + month	Firm + month
Wald p (post3 = post1)	0.213	0.495	0.315	0.379	0.554	0.362
N	47,226	47,226	47,226	47,226	47,226	47,226
R-squared	0.17	0.17	0.17	0.17	0.17	0.17

Notes. Each column regresses monthly abnormal returns on the exposure measure named in the rows interacted with three post-period bins, with firm and month fixed effects, on the January 2018 to December 2025 monthly panel of Section 5. The omitted category is the January 2018 to November 2022 pre-period, so coefficients are relative to the whole pre-period average rather than to the first half of 2022 as in Figure 4, and the two exhibits are not comparable coefficient by coefficient. Columns (1)–(3) use the high-exposure indicators and columns (4)–(6) the standardized continuous exposures. Standard errors are clustered two-way on firm and month. The Wald row reports the p-value for equality of the 2025 and the 2022m12–2023m12 interactions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.13: Nesting test: technical exposure versus its deployable and marked-down components

	(1)	(2)	(3)	(4)
Technical exposure	-0.001	-0.032		
	(0.038)	(0.026)		
Deployable exposure			0.151	0.009
			(0.063)	(0.065)
Institutional markdown			-0.210	-0.096
			(0.048)	(0.096)
Wild-cluster p: Technical = 0	0.977	0.198		
Wild-cluster p: Deployable = Markdown			0.042	0.598
Industry FE	Yes	Yes	Yes	Yes
Firm controls	No	Yes	No	Yes
N	489	404	489	404
R-squared	0.17	0.29	0.20	0.29

Notes. Continuous counterpart of Table 3. The dependent variable is the market-adjusted cumulative abnormal return of equation (7) over December 2022 to December 2025, winsorized at the 1st and 99th percentiles, on the balanced sample. Columns (1) and (2) enter technical exposure alone. Columns (3) and (4) enter deployable exposure and the institutional markdown together, which is the decomposition implied by $\text{Tech}^F = \text{Inst}^F + \Delta^{F,AC}$. All three components are divided by a single common factor, the standard deviation of technical exposure in the balanced cross-section (0.028 on the 0–1 exposure scale), so that the identity is preserved and coefficients read as cumulative abnormal return per one technical-exposure standard deviation of the component. Standardizing the components separately would rescale them by different factors, after which equality of the two coefficients would no longer represent the nesting restriction. All columns include pre-event finance-subsector fixed effects, and parentheses report standard errors clustered on the same subsector. The two wild-cluster rows report bootstrap p-values based on 9,999 replications, for the hypothesis that technical exposure earns nothing in columns (1) and (2) and for the nesting restriction that deployable exposure and the markdown earn the same coefficient in columns (3) and (4). No significance stars are reported.

Table A.14: Alternative long-horizon return measures

	BHAR	BHAR	BHAR	RW	RW	RW
High technical exposure	-0.145			-0.086		
	(0.111)			(0.066)		
High deployable exposure		-0.149			-0.088	
		(0.192)			(0.114)	
High markdown			-0.338			-0.200
			(0.137)			(0.081)
Outcome	BHAR	BHAR	BHAR	Rel. wealth	Rel. wealth	Rel. wealth
Wild-cluster p-value	0.214	0.527	0.184	0.214	0.527	0.184
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes
N	404	404	404	404	404	404
R-squared	0.23	0.23	0.23	0.23	0.23	0.23

Notes. Same balanced sample, pre-event finance-subsector fixed effects, frozen October 2022 controls, and clustering as Table 3, with the headline cumulative abnormal return replaced by the two conventional compounded measures. Columns (1)–(3) use the buy-and-hold abnormal return, $BHAR_i = \prod_t(1+r_{it}) - \prod_t(1+r_{mt})$. Columns (4)–(6) use relative wealth, $RW_i = \prod_t(1+r_{it}) / \prod_t(1+r_{mt}) - 1$. Both outcomes are winsorized at the 1st and 99th percentiles on the balanced cross-section. The two measures are not independent tests of the same hypothesis. On a balanced window every firm holds the market index over the same months, so the market wealth path is a single constant and the two measures are affine transformations of the same terminal firm wealth, which is why they return identical wild-cluster p-values. They are reported because the compounded measures are much more sensitive to the right tail of three-year returns than the additive cumulative abnormal return is, so the choice of headline measure should be visible rather than implicit. The wild-cluster row reports bootstrap p-values based on 9,999 replications. No significance stars are reported.

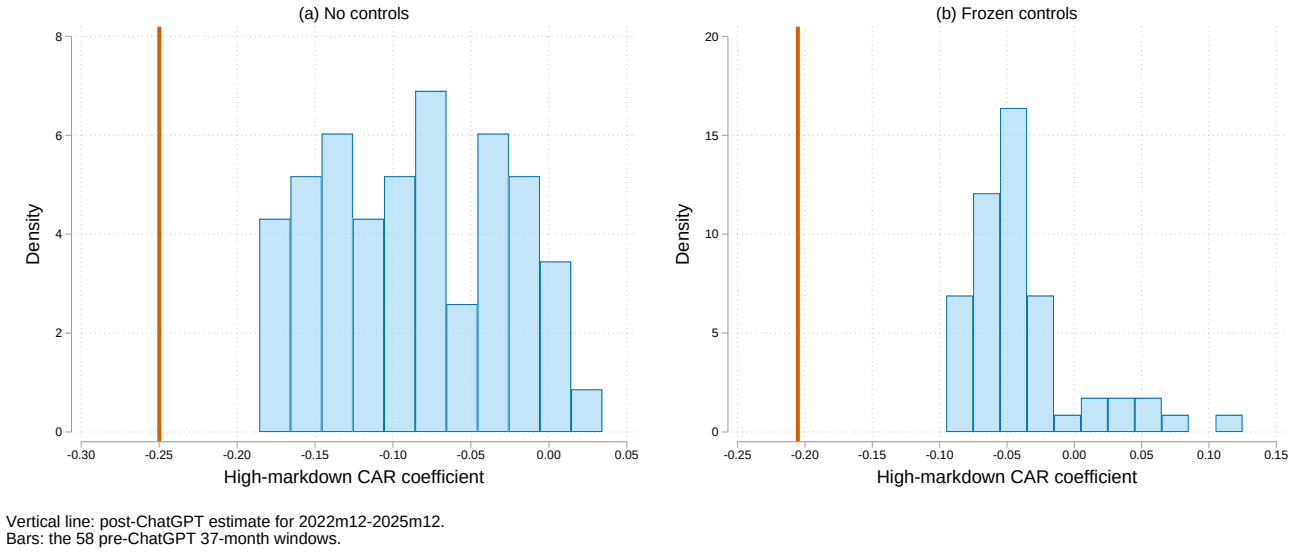


Figure A.1: Rolling placebo distribution of the high-markdown cumulative-return differential. Histograms show the coefficient on the high-markdown indicator across the 58 pre-ChatGPT thirty-seven-month windows described in Table A.15, without firm controls in panel (a) and with the frozen October 2022 controls in panel (b). The vertical line marks the post-ChatGPT estimate for the December 2022 to December 2025 window. In both panels the post-ChatGPT estimate lies below every pre-ChatGPT window estimate. Because the windows overlap, the position of the line is an empirical benchmark rather than a p-value.

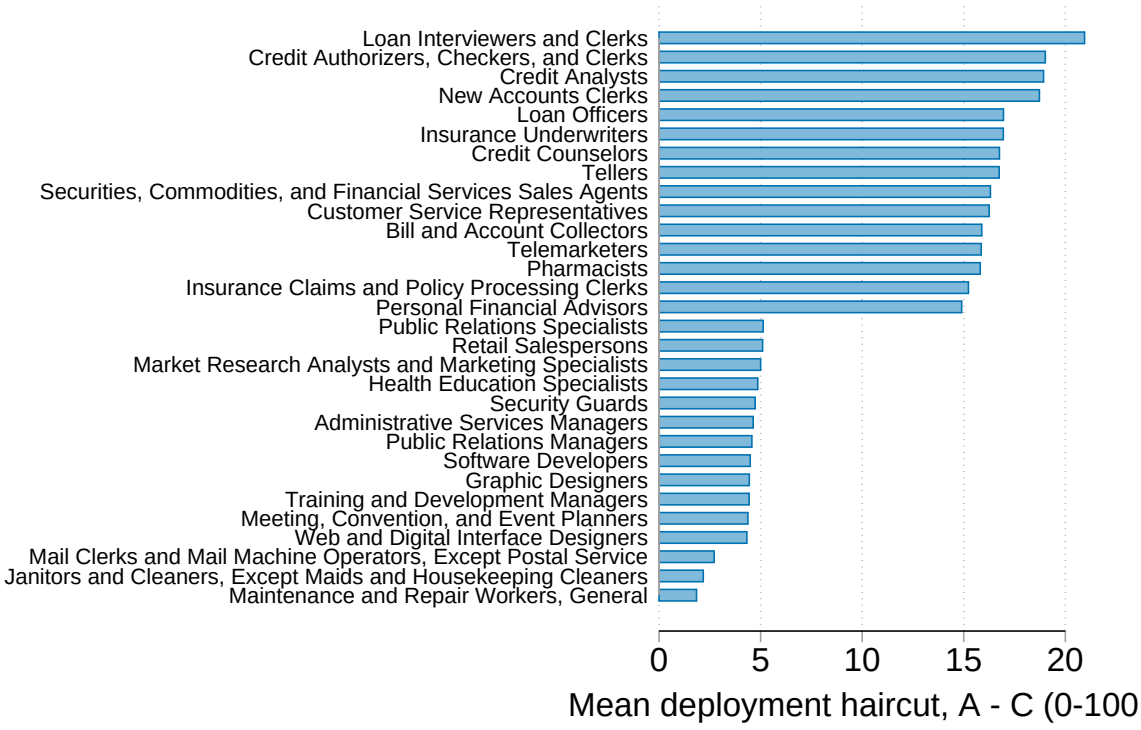


Figure A.2: Institutional markdown by occupation, top and bottom fifteen. Bars are mean institutional markdowns $A - C$ (0–100 points), averaged over models and tasks. This is the descriptive counterpart of Figure 5, shown without confidence intervals.

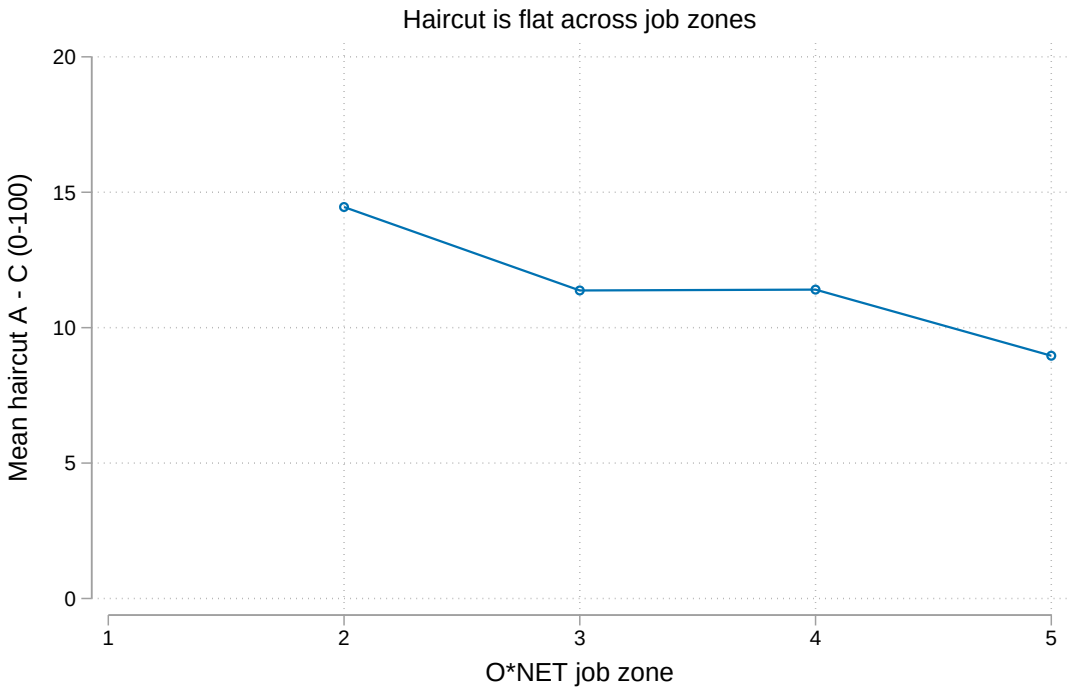


Figure A.3: Institutional markdown across O*NET job zones. Bars report mean employment-weighted institutional markdowns $A-C$ (0–100 points) by O*NET job zone, where 1 indicates little preparation and 5 indicates extensive preparation. The markdown is roughly flat across job zones, which helps separate institutional co-dependence from education or preparation requirements.

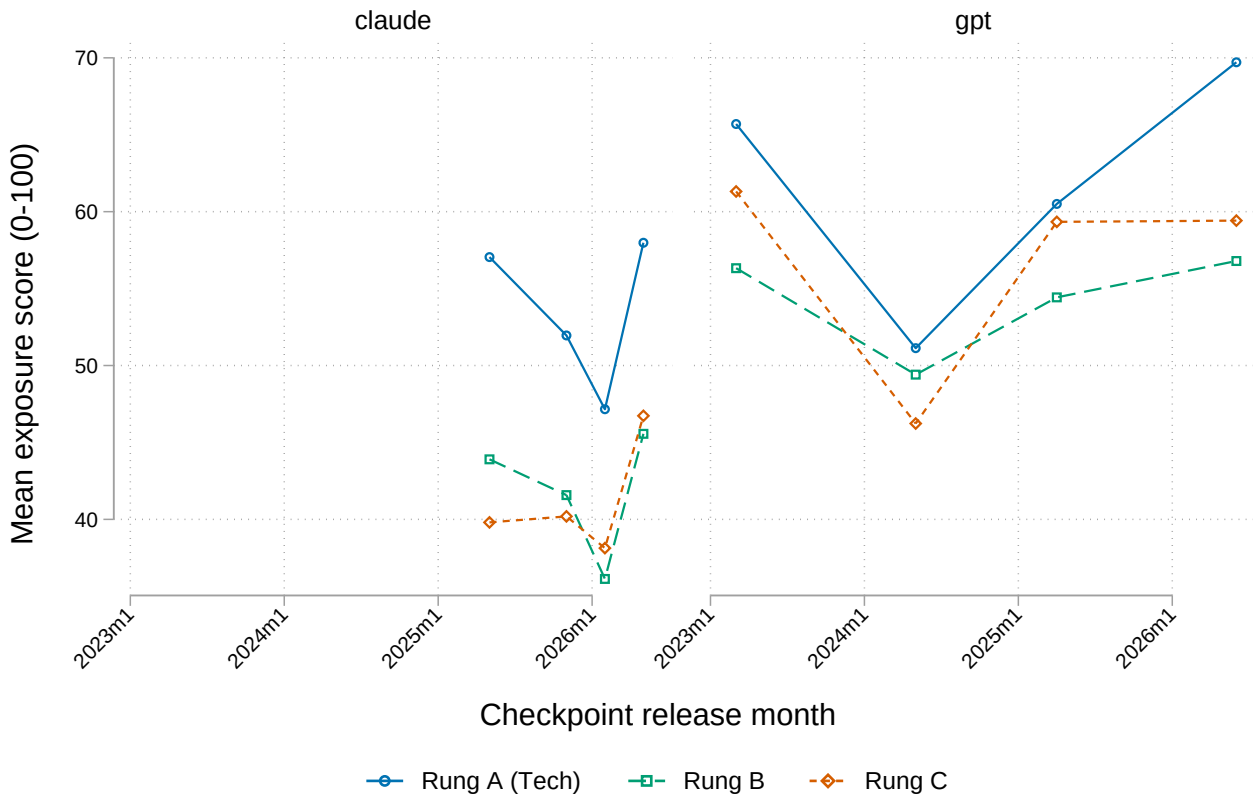


Figure A.4: Exposure trajectory across model checkpoints. Mean exposure (0–100) at rungs A, B, and C by checkpoint release month, plotted separately for each developer family. Because the scoring rubric's "current frontier" anchor is indexical, the level trajectories mix true capability change with rater-calibration drift and should be read as description.

Table A.15: Rolling placebo distribution of the high-markdown differential

	No controls	Frozen controls
Post-ChatGPT window, 2022m12–2025m12	-0.250	-0.205
Distribution over 58 pre-ChatGPT 37-month windows		
Minimum	-0.186	-0.095
10th percentile	-0.162	-0.083
Median	-0.083	-0.048
90th percentile	-0.006	0.041
Maximum	0.017	0.106
Standard deviation	0.057	0.042
Placebo windows at least as negative	0 of 58	0 of 58

Notes. Each entry is the coefficient on the high-markdown indicator from the specification of Table 3, estimated over a thirty-seven-month window. The first row uses the post-ChatGPT window. The distribution rows summarize the 58 pre-ChatGPT windows whose start months run from January 2015 to October 2019, so that every placebo window closes on or before October 2022, the last pre-event month. Sample definition, median-split exposure sorts, pre-event finance-subsector fixed effects, the 1st and 99th percentile winsorization applied within each window, and the frozen October 2022 controls are identical across windows, and the post-ChatGPT window is estimated through the same code path. Windows overlap by up to thirty-six months and are therefore strongly dependent, so the count in the final row is an empirical placebo benchmark and not a size-controlled p-value. Exposure and the frozen controls are measured in 2022 and therefore after most placebo windows close, which is inherent to holding the treatment variable fixed across windows.

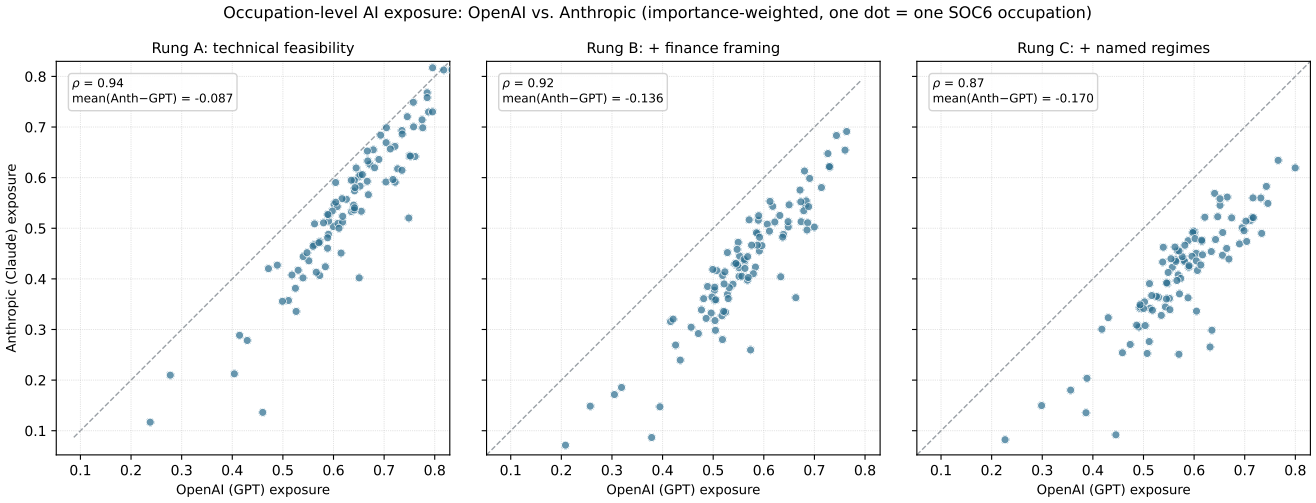


Figure A.5: Occupation-level exposure, GPT versus Claude family, by prompt rung. Each dot is one of the 99 finance-and-insurance occupations, plotting the importance-weighted mean exposure from the four GPT-family checkpoints (horizontal axis) against the importance-weighted mean from the four Claude-family checkpoints (vertical axis) at each rung. All but one occupation lie below the 45-degree line at rung A, and every occupation does at rungs B and C. The mean gap between the families widens from 8.7 points at rung A to 17.0 points at rung C (0–100 points). The cross-vendor Spearman rank correlation is 0.93, 0.90, and 0.85 at rungs A, B, and C. The figure is descriptive. Points are importance-weighted occupation means without clustered inference.

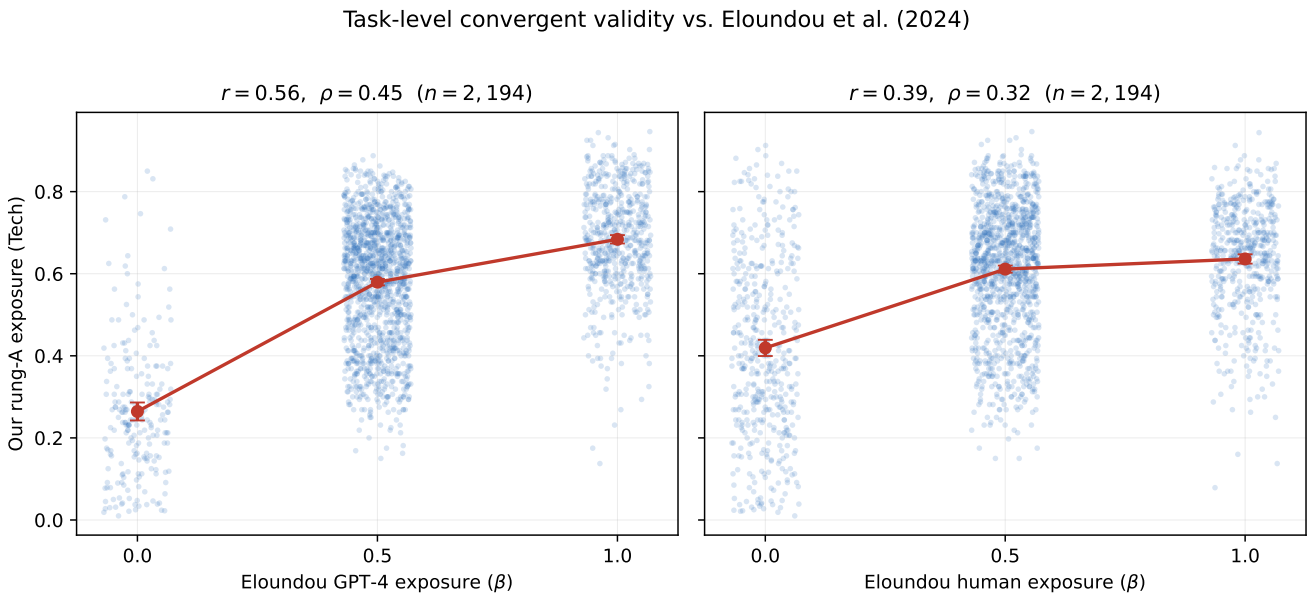


Figure A.6: Task-level convergent validity against Eloundou et al. (2024). Each light point is one of the 2,194 merged tasks, plotting our rung-A technical-feasibility score (vertical axis) against the Eloundou β exposure rating (horizontal axis), using the GPT-4 rating in the left panel and the human-annotator rating in the right panel. The solid markers trace the mean rung-A score within each rating bin. At the task level, the rung-A score correlates 0.56 with the GPT-4 rating and 0.39 with the human annotation. The $A - C$ markdown correlates only 0.21 with the GPT-4 rating at the task level, consistent with the markdown carrying institutional content that the feasibility benchmark omits by design.

References

Andrews, I. and Farboodi, M. (2025). Do markets believe in transformative AI? NBER Working Paper 34243, National Bureau of Economic Research. Revised July 2026.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2024a). Artificial intelligence and firms' systematic risk. Working paper, SSRN 4868770.

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

Borri, N., Tsyvinski, A., and Liu, Y. (2026). AI premium. NBER Working Paper 35451, National Bureau of Economic Research.

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

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47.

Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E. J., and Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work. IMF Staff Discussion Note 2024/001, International Monetary Fund.

Dell'Acqua, F., McFowland, Edward, I., Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker

productivity and quality. *Organization Science*, 37(2):403–423. Earlier circulated as Harvard Business School Working Paper No. 24-013 (2023).

Eisfeldt, A. L., Schubert, G., and Zhang, M. B. (2023). Generative AI and firm values. NBER Working Paper 31222, National Bureau of Economic Research.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308. Earlier circulated as arXiv:2303.10130 (2023).

Espinal Maya, C. (2026). Measuring what cannot be surveyed: LLMs as instruments for latent cognitive variables in labor economics. Working Paper arXiv:2604.02403, arXiv.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Felten, E. W., Raj, M., and Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Gmyrek, P., Berg, J., and Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. ILO Working Paper 96, International Labour Organization.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. (2025). Which economic tasks are performed with AI? evidence from millions of Claude conversations. Anthropic economic index report, Anthropic.

Lindenlaub, I., Oh, R., Rodriguez, M. A., and Veldkamp, L. (2026). Beyond exposure: Predicting AI adoption based on comparative advantage. NBER Working Paper 35271, National Bureau of Economic Research.

Makridis, C. A. (2026). The organizational transmission of AI: The role of managers on AI adoption and impact. CESifo Working Paper 12373, CESifo.

Makridis, C. A., Brynjolfsson, E., and Kazinnik, S. (2026). Ai (at work) in finance. *SSRN working paper*.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Parshall, D. and Lopez-Luzuriaga, A. (2026). Measuring AI's economic reach: A multi-dimensional task taxonomy. Working Paper 2026-005, The George Washington University, Center for Economic Research.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot.

Revelio Labs (2026). Workforce data platform: Profiles, job postings, transitions, sentiment, and layoffs. Company data documentation. Accessed July 11, 2026.

Webb, M. (2020). The impact of artificial intelligence on the labor market. Job market paper, Stanford University. Available at SSRN 3482150 (first posted November 2019).

Yin, M., Vu, H., and Persico, C. (2026). How (un)stable are LLM occupational exposure scores? evidence from multi-model replication. NBER Working Paper 35110, National Bureau of Economic Research. Also available as SSRN 6656303.

Yu, Y., Fleming, M., Hampton, L., Combemale, C., and Thompson, N. (2026). AI adoption in S&P 500 firms.