Capable but Not Deployable: Separating AI Feasibility from Institutional Co-Dependence in Finance

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

August 13, 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. In regulated industries, deployable exposure rather than technical feasibility is the economically relevant measure of AI exposure.

Keywords: AI exposure; large language models; occupational tasks; finance labor markets; technology adoption; measurement.

JEL Codes: C52; C81; 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; Lisa 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 full responsibility for the content of the publication. These views are our own and do not reflect those of any affiliated institutions.

1 Introduction

A finance task can be technically easy for a frontier model and still be difficult for a financial institution to deploy. 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. It is 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 paper measures the wedge between those two objects and shows that it is large, unanimous across frontier models, and concentrated in exactly the occupations where financial regulation binds.

The standard empirical approach to AI exposure starts from task content. Researchers score O*NET task statements for the extent to which AI can reduce the time required to complete them, increasingly by using large language models as task raters (Eloundou et al., 2024), building on earlier occupation- and technology-linkage measures (Felten et al., 2021; Webb, 2020). These scores describe the technological frontier. They are less informative about deployment in settings where the task boundary depends on institutional structure. Finance is such a setting: a credit analyst, underwriter, broker, adviser, compliance officer, accountant, or customer-service representative does not simply produce text or analysis, but operates inside a regulated production system. Feasibility-based exposure scores are nonetheless carried into finance applications, including firm-level exposure measures used as asset-pricing regressors (Eisfeldt et al., 2023). If cash flows respond to what firms can deploy rather than to what models can do, feasibility-based exposure mismeasures the priced object, and mismeasures it most severely for the regulated firms that

constitute the finance sector.

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,[1] 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.

The measurement design is a prompt ladder. We score 2,199 O*NET task statements across 99 finance-and-insurance occupations with eight frontier model checkpoints from the GPT and Claude families. Rung A asks for bare technical feasibility. Rung B adds that the task is performed inside a financial institution. Rung C names the relevant finance governance domains. We measure the $A - C$ difference, which we call the institutional markdown, or deployment haircut. Keeping the task, model, score scale, and output schema fixed across rungs lets us read the markdown from within-model differences rather than from cross-model levels. Within-model comparisons remove the calibration confound that arises because one model may use the same numerical scale more aggressively than another.

The main result is that the institutional markdown is large. The employment-weighted mark-

[1]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.

down 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. 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. 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.[2]

Institutional context does not only lower measured exposure. It also makes exposure harder for models to agree on: 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 therefore 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 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.

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

Our contribution is a deployability-adjusted measure of AI exposure for the finance industry. The exposure-measurement literature builds on task-based measures of machine-learning suitability, patent-task overlap, and AI occupational linkage (Brynjolfsson et al., 2018; Webb, 2020; Felten et al., 2021), and has moved toward GPT- and LLM-based measures of task exposure (Eloundou et al., 2024; Eisfeldt et al., 2023; Gmyrek et al., 2023; Cazzaniga et al., 2024; Handa et al., 2025). Motivated by the surge in AI use in finance (Makridis et al., 2026), we depart from this literature in three ways. Rather than treating model feasibility as the exposure measure, we estimate both bare technical feasibility and the markdown that finance context and regulatory requirements impose on it, holding the task, model, and score scale fixed so that the markdown is identified within model. Rather than relying on a single rater at a single point in time, we score every task across eight frontier model checkpoints from two independently developed model families, which lets exposure vary with the capability frontier and lets us measure where models agree and where they diverge. And rather than treating institutional friction as the generic organizational drag that existing work shows slows AI adoption (Makridis, 2025), we anchor the friction layer to named regulatory regimes and build it into the exposure measure itself, at the task level, before any labor-market or return outcome is examined. Implementing the approach on the NAICS 52 task universe shows that the markdown is one-fifth of measured feasibility on average, is concentrated in regulated, client-facing occupations, and is not an artifact of task length or job-zone complexity.

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 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. The universe is broader than a list of finance-titled occupations. It includes occupations that financial institutions actually employ—for example, customer-service representatives, software developers, marketing specialists, compliance officers, accountants, insurance claims workers, tellers, under-

writers, loan officers, and credit analysts. Deployability depends on where the task is performed, not only on the occupation title, so the industry-employment universe is the relevant unit for the economic question.

The finance-industry restriction uses occupational employment in NAICS 52. We keep SOC occupations that clear the finance-employment floor and then merge their O*NET task statements, task importance, task frequency, and job-zone measures. The task-level file therefore contains three types of information before any model is queried: the task text, the occupation and finance-industry employment weights, and pre-scoring task controls such as 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. The empirical object therefore relies 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.

Other adoption benchmarks—the Anthropic Economic Index, 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 are narrower by design: they establish that the measurement instrument produces a stable and interpretable finance markdown before asking whether that markdown predicts realized adoption, employment, wages, or returns.

2.5 Current limitations

We note 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. Aggregating the occupation-level measure to individual firms requires firm-by-occupation staffing data and is left to future work.

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 cor-

rect 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 quarantined diagnostic proxy built from the rung-C model responses, but that proxy is not a paper-grade co-dependence measure because it is derived from the same model outputs used to score the constrained rung. The main evidence in this version is therefore H-free: 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

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

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, and the cross-model agreement pattern. 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.

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.5 shows the task-level scatter). Figure 1 shows the occupation-level 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 behaves differently. It correlates only 0.23 with the Eloundou GPT-4 rating

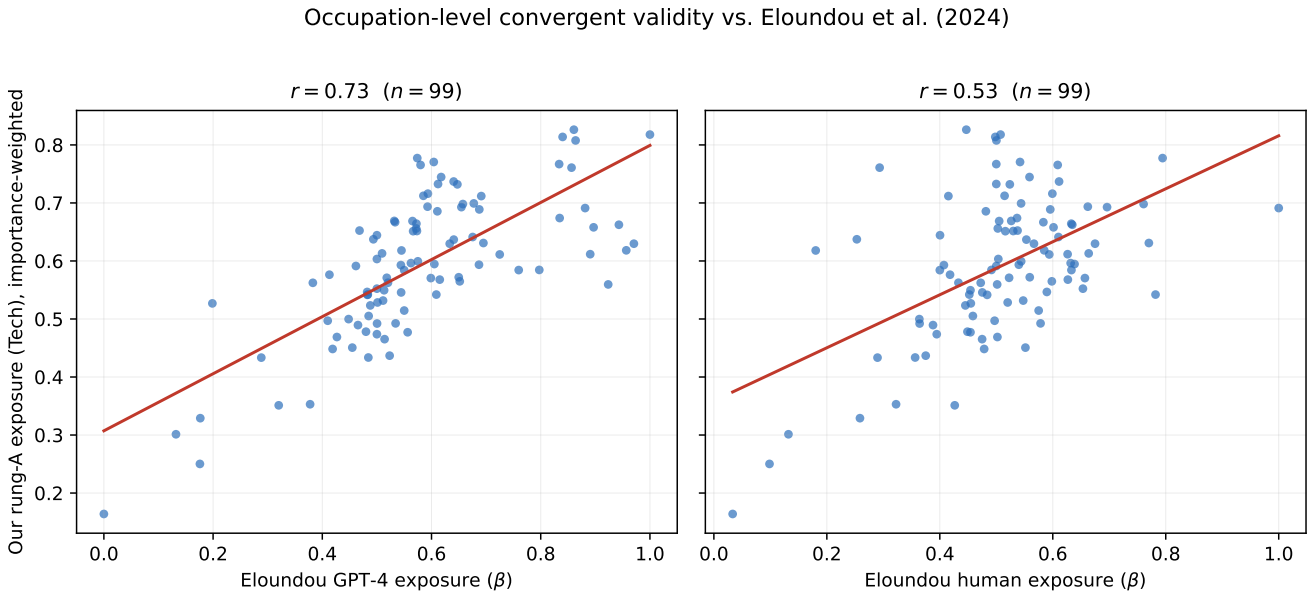


Figure 1: Occupation-level convergent validity against Eloundou et al. (2024). Each point is one of the 99 finance-and-insurance occupations. The vertical axis is our importance-weighted rung-A technical-feasibility score. The horizontal axis is the Eloundou β exposure aggregated over the identical tasks 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.

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 1 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 2 reports the paper-grade employment-weighted estimates 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, not a calibration artifact.

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 4 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 3 lists the fifteen largest and fifteen smallest occupation-level markdowns, and Figure 2 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 maintenance 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 3 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. 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.

Two artifact checks support this interpretation. Appendix Table A.3 shows that task word count does not predict the markdown once model fixed effects and job-zone dummies are included. If anything, longer tasks receive lower technical scores. Appendix Figure A.2 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 4 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 5 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.4 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.

The appendix reports two additional diagnostic families. 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. The paper-grade version replaces it with an independent constraint vector and should be evaluated separately.

4.5 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.[3] 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 the 1,907 surveyed task statements rather than on the full universe, because the per-call rationale text for the 292 later-admitted statements is not held on the analysis machine.

[3] 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.

4.6 Model vintage evidence is descriptive

Appendix Table A.8 and Appendix Figure A.3 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 best read as a caution against overinterpreting cross-vintage levels, not as evidence that capability is flat over time.

4.7 What the current results establish

The current results establish four 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. These are measurement results, and they are the object the paper delivers. How the deployable component converts into firm-level exposure variables, and whether markets price it, are questions the measure makes askable but that we leave to future work.

5 Conclusion

This paper separates what frontier models can do to finance tasks from what financial institutions can deploy. The gap between the two objects is large. 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 contribution is a measurement object, not only a new score. 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. Aggregating the occupation-level decomposition to firms with firm-by-occupation staffing shares, and testing whether markets price the deployable component around AI capability shocks and in the cross-section of expected returns, is the natural next step and is left to future work.

Two extensions follow directly. Task-specific constraint codes, assigned under the codebook in Appendix A, will convert the governance-context markdown into a full deployability score and allow dimension-specific tests of whether the right constraints bind for the right task families. An independently coded co-dependence vector will also discipline the measurement-error interpretation: if cross-model disagreement at the bare-capability rung loads on coded institutional co-dependence, the instability documented in recent work is a symptom of an omitted economic layer rather than noise to be averaged away.

Tables and Figures

Table 1: 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.

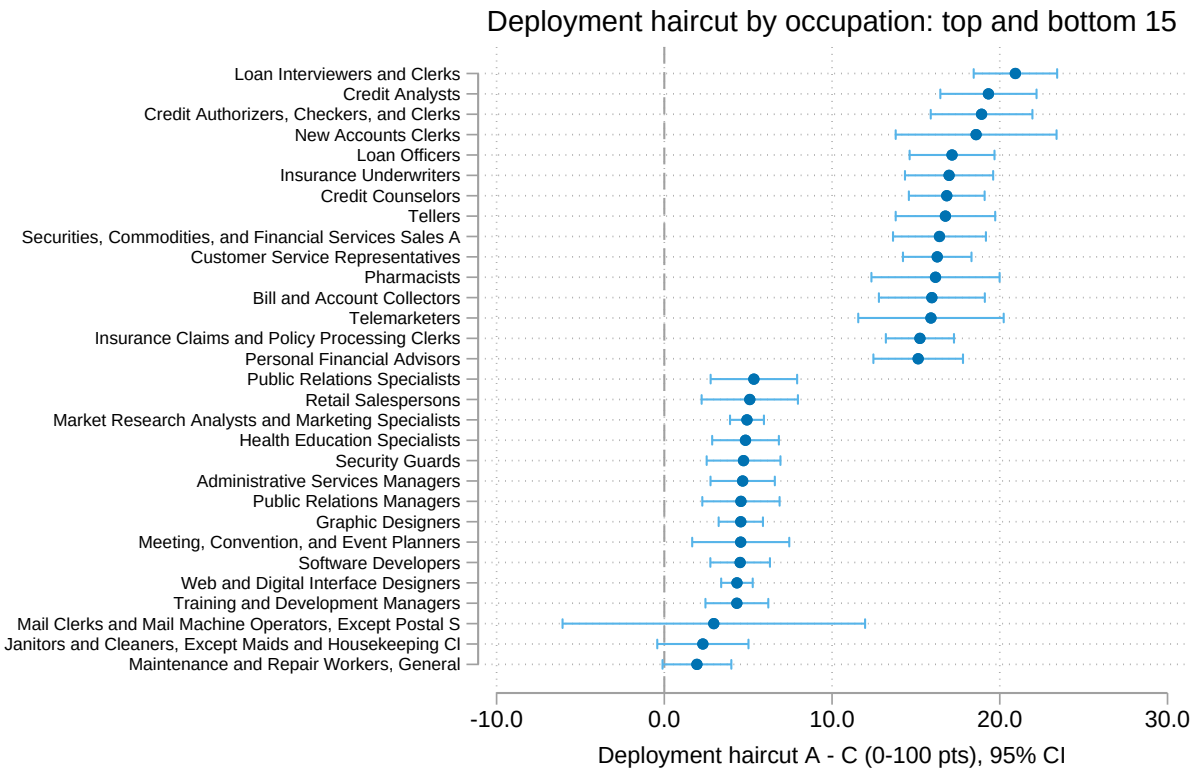


Figure 2: Institutional markdown by occupation, with 95% confidence intervals. Each row is one of the fifteen most- and fifteen least-affected six-digit SOC occupations, ordered by the pooled institutional markdown $A - C$ (0–100 points). Points are importance-weighted markdowns averaged over the eight 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.

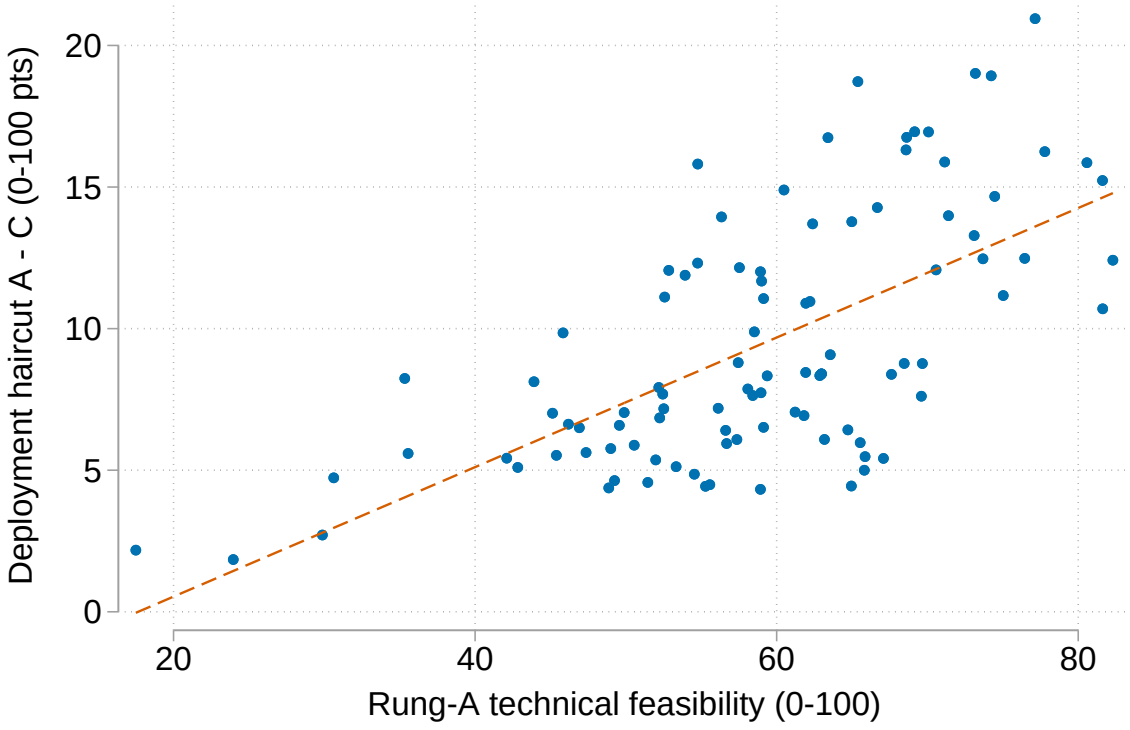


Figure 3: The institutional markdown against technical feasibility across occupations. Each point is one of the 99 finance-and-insurance occupations, plotting the mean institutional markdown $A - C$ (vertical axis, 0–100 points) against mean rung-A technical feasibility (horizontal axis, 0–100 points). 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 2: 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 $B - C$.

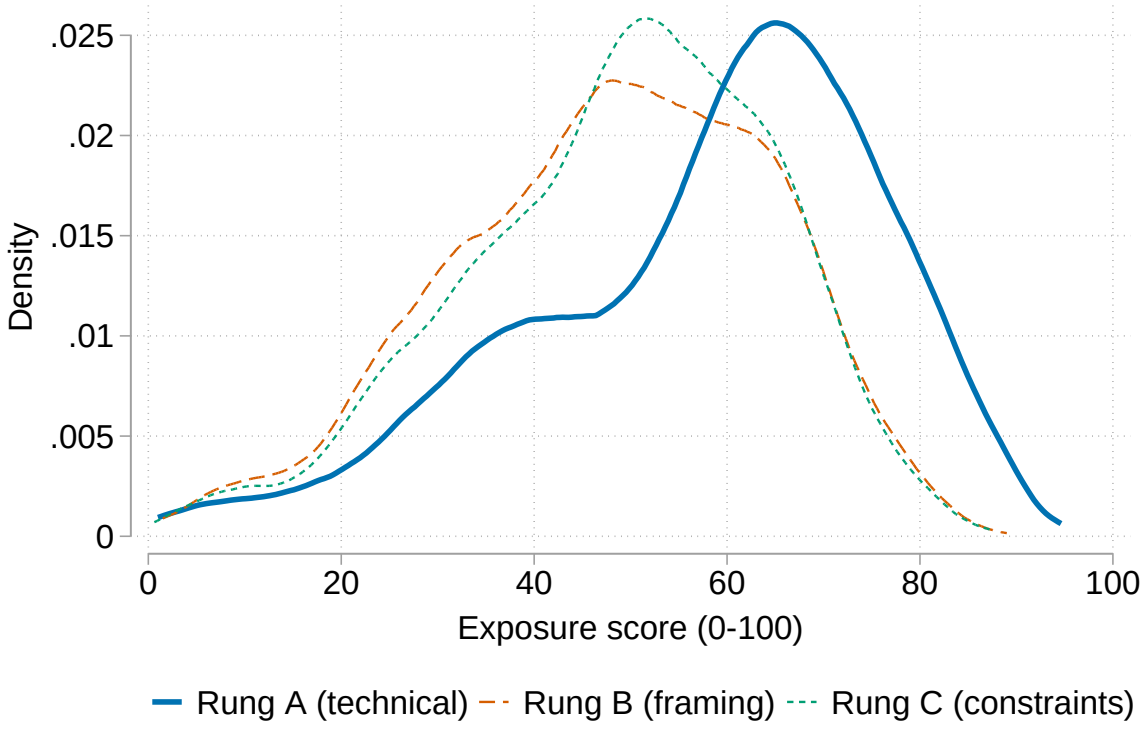


Figure 4: Task-level distribution of exposure scores by prompt rung. Kernel densities of task-mean exposure scores (0–100), pooled across models, at rung A (technical feasibility), rung B (finance framing), and rung C (named governance regimes). Adding finance context shifts the distribution left, which is the distributional signature of the institutional markdown.

Table 3: 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 2 plots the same comparison with confidence intervals.

Table 4: 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.

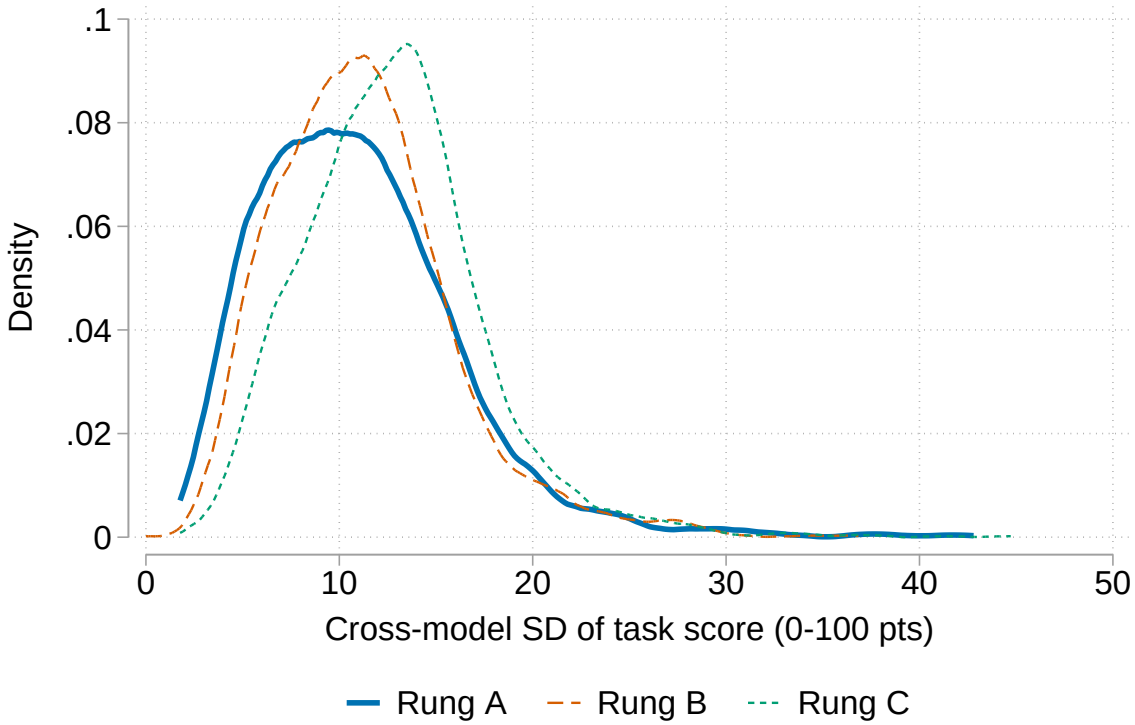


Figure 5: Cross-model dispersion by prompt rung. Kernel densities of each task's cross-model standard deviation (0–100 points) at rungs A, B, and C. Disagreement widens as institutional context is added, consistent with the view that models agree more on bare capability than on deployability inside finance.

A Coding Guide for Finance-Industry Institutional Constraints

This appendix defines the five institutional-constraint dimensions that constitute 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 and adjudication protocol

The post-pivot task universe contains 2,199 tasks, so the original two-author full hand-coding protocol is no longer the main production plan. The paper-grade H_k^F vector will be built in two steps. First, the full task universe is coded using a structured coding instrument that applies the five dimensions above. Second, the authors audit and adjudicate a fresh random subset, with oversampling of high-markdown and borderline tasks, to estimate reliability and correct systematic coding errors. The two legacy author-coded files from the 360-task universe are retained as an internal benchmark, not as the gold-standard code for the current universe.

For the audited subset, each author scores the task independently before observing the other's code, the exposure scores, or the return outcomes. Disagreements are resolved by re-reading the task against the anchors and boundary rules in this appendix. Where ambiguity remains, the pre-specified tie-break is the conservative rule in Section A.1: assign the lower code unless the occupation and task family clearly imply the stronger one. The replication package reports pre-adjudication and post-adjudication codes, weighted kappa for each ordinal dimension, an intraclass correlation for H_k^F, and all primary results under both the production vector and the adjudicated audit 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. Task length does not predict the markdown. Longer tasks 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.3 is the figure counterpart. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

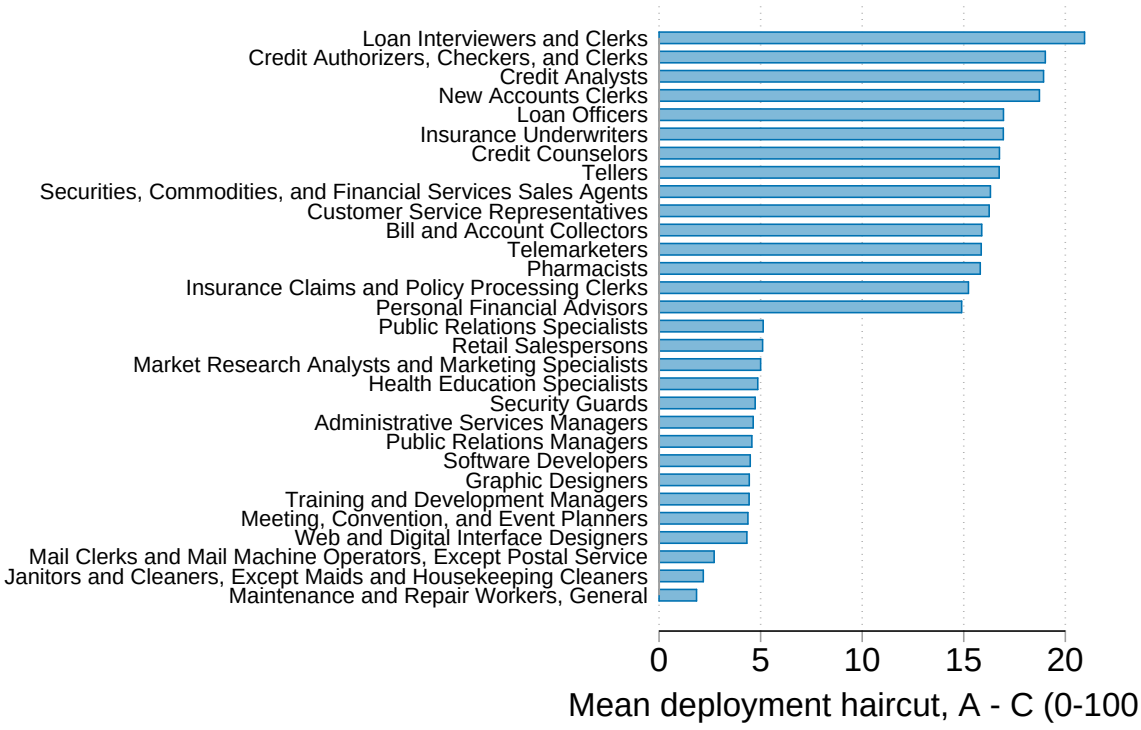


Figure A.1: Institutional markdown by occupation, top and bottom fifteen. Bars are mean institutional markdowns $A - C$ (0–100 points), averaged over models and tasks. This is the descriptive counterpart of Figure 2, shown without confidence intervals.

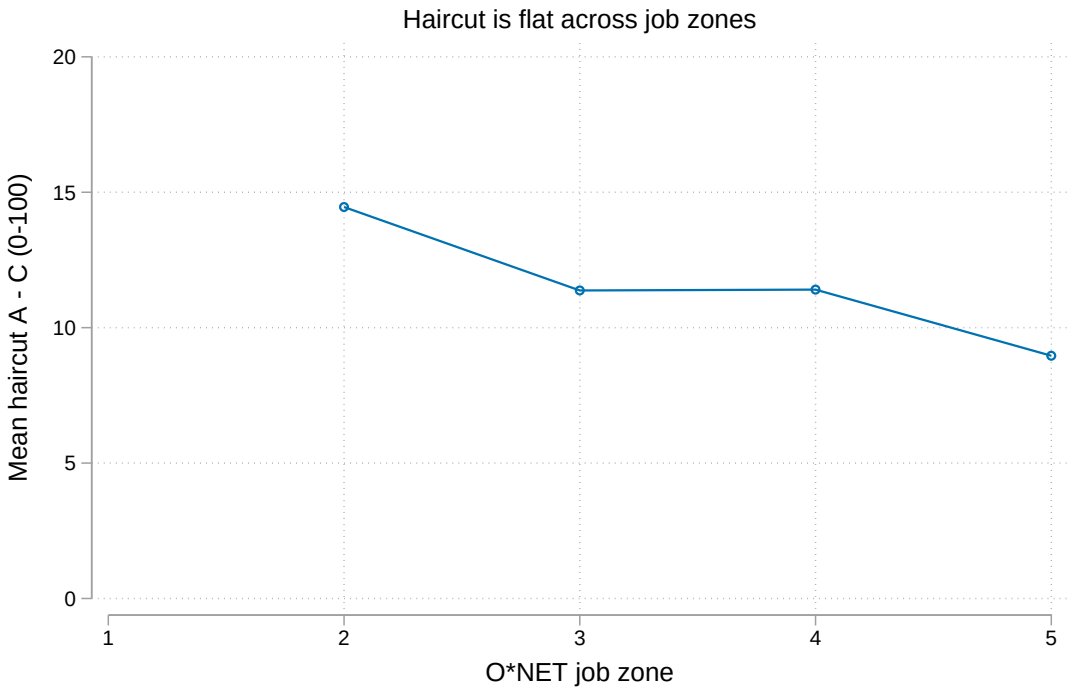


Figure A.2: Institutional markdown across O*NET job zones. Bars report mean employment-weighted institutional markdowns $A-C$ (0–100 points) by O*NET job zone, where 1 indicates little preparation and 5 indicates extensive preparation. The markdown is roughly flat across job zones, which helps separate institutional co-dependence from education or preparation requirements.

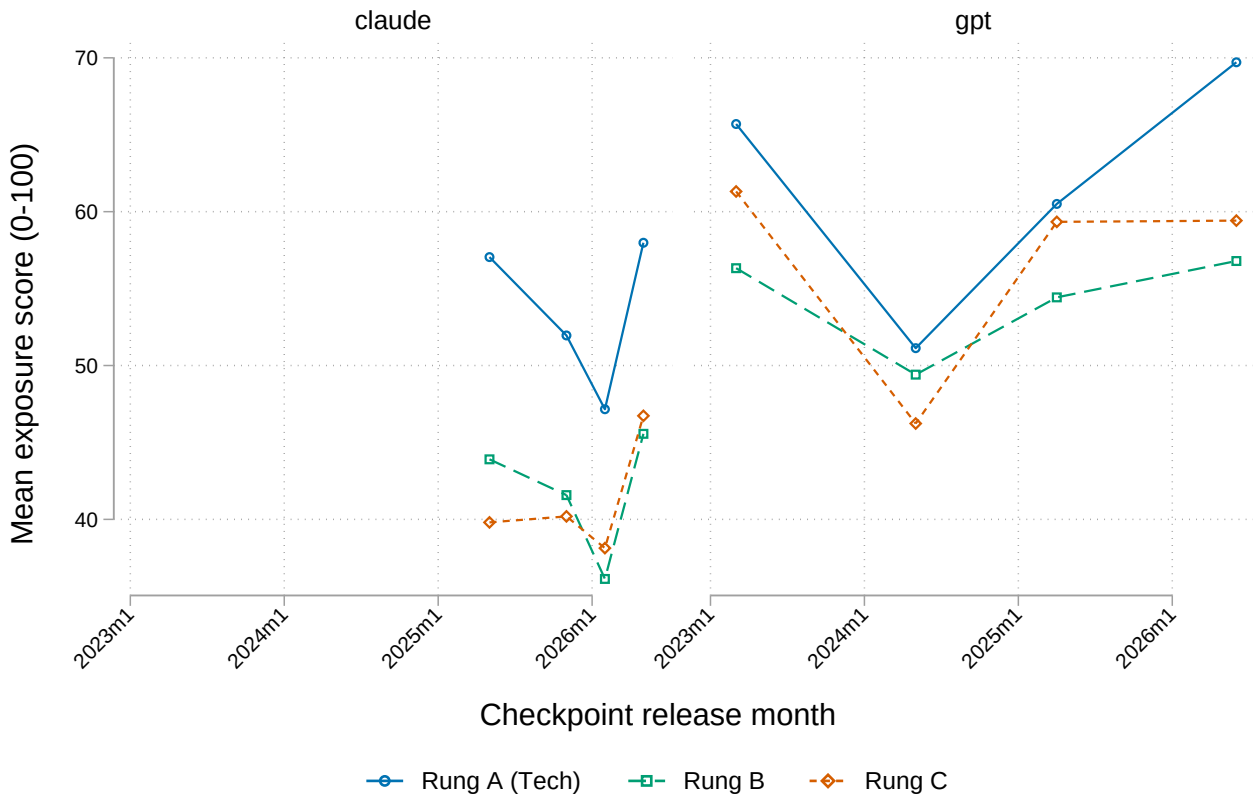


Figure A.3: Exposure trajectory across model checkpoints. Mean exposure (0–100) at rungs A, B, and C by checkpoint release month, plotted separately for each developer family. Because the scoring rubric's "current frontier" anchor is indexical, the level trajectories mix true capability change with rater-calibration drift and should be read as description.

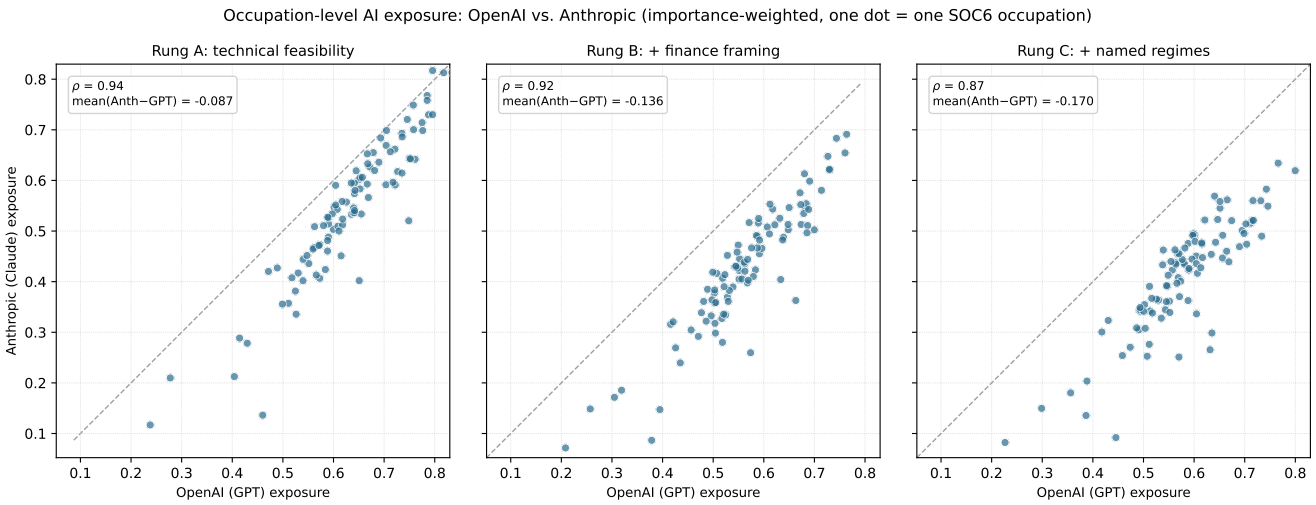


Figure A.4: Occupation-level exposure, GPT versus Claude family, by prompt rung. Each dot is one of the 99 finance-and-insurance occupations, plotting the importance-weighted mean exposure from the four GPT-family checkpoints (horizontal axis) against the importance-weighted mean from the four Claude-family 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.

References

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.

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

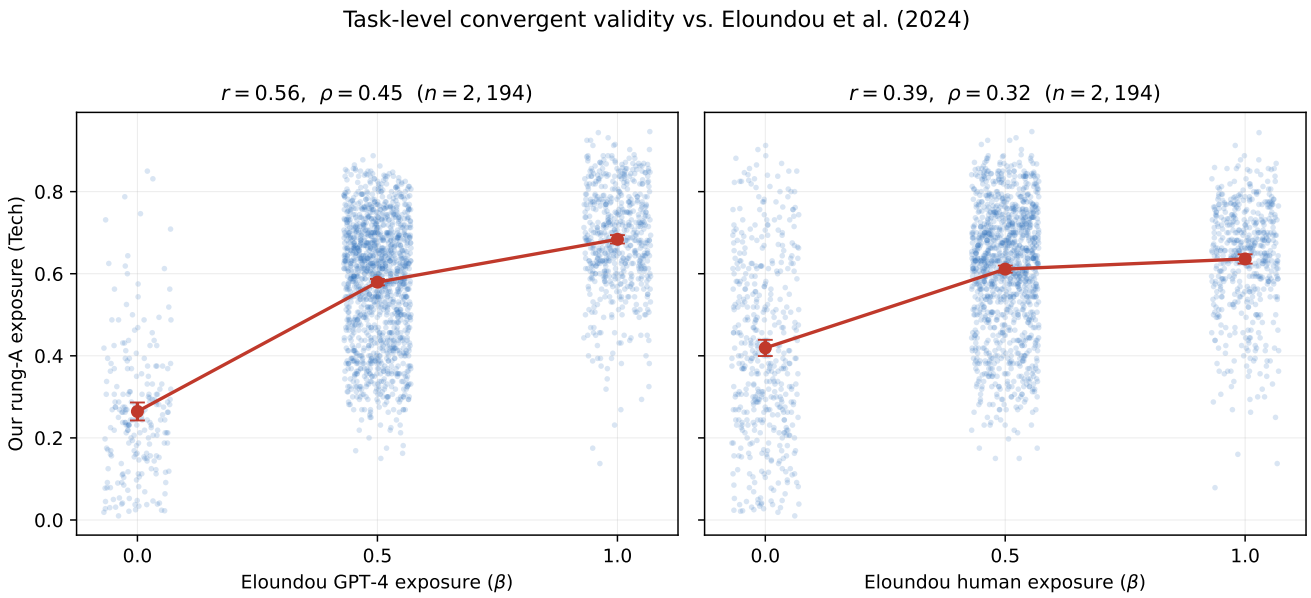


Figure A.5: Task-level convergent validity against Eloundou et al. (2024). Each light point is one of the 1,902 merged tasks, plotting our rung-A technical-feasibility score (vertical axis) against the Eloundou β exposure rating (horizontal axis), using the GPT-4 rating in the left panel and the human-annotator 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 0.58 with the GPT-4 rating and 0.39 with the human annotation. The $A - C$ markdown correlates only 0.23 with the GPT-4 rating at the task level, consistent with the markdown carrying institutional content that the feasibility benchmark omits by design.

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.

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.

Makridis, C. A. (2025). The organizational transmission of AI: The role of managers on AI adoption and impact. CESifo Working Paper 12373, CESifo. Available at SSRN 5401053.

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

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.