A Protocol for Rater-Invariant Measures from Language Models With an Application to AI Exposure

Benjamin Verschuere · Angus Cameron
Liminal Capital

Conference on Frontiers in Machine Learning and Economics

Federal Reserve Bank of Philadelphia and Chicago Booth Center for Applied AI · November 6, 2026

Disclosure: both authors are employed by Liminal Capital LLC, a CFTC-registered Commodity Trading Advisor and NFA member (NFA ID 0525645). Views are the authors' own. The analysis uses only public data and public model APIs; nothing here is investment advice. Working paper: SSRN 7114299.

Disclosure

This paper is an independent academic working paper prepared by the authors solely in their research capacity. It is submitted for academic distribution and represents the authors' independent contribution to the literature on the measurement of AI exposure. This paper does not constitute investment advice, a trading recommendation, or a solicitation of any kind. Nothing in this paper should be construed as relating to any commodity interest account, agreement, or transaction, nor as an offer or invitation to invest in any fund, managed account, or trading program. It makes no reference to the past or projected performance of any trading program, commodity pool, or managed account, and contains no hypothetical performance results within the meaning of applicable commodity trading regulations. The views and conclusions expressed herein are solely those of the authors in their individual research capacity and do not represent the views or positions of Liminal Capital LLC, its principals, employees, affiliates, or clients. Institutional affiliations are disclosed for identification purposes only. The authors are employed by Liminal Capital LLC, a CFTC-registered Commodity Trading Advisor and NFA Member (NFA ID: 0525645). This relationship is disclosed in accordance with standard academic conflict of interest practice and does not imply that this paper was prepared in connection with Liminal Capital's commercial activities. All external data sources used in the analysis are publicly available: the O*NET task, task-ratings, and abilities databases (National Center for O*NET Development); the BLS Occupational Employment and Wage Statistics (OEWS) for employment weights; the AI Occupational Exposure index of Felten, Raj, and Seamans; the Dingel and Neiman teleworkability classification; and firm-reported AI adoption from the U.S. Census Bureau Business Trends and Outlook Survey (BTOS). The task-level exposure scores analyzed here were generated by the authors by applying a fixed public scoring rubric to the O*NET task set through the published APIs of six large language models; no proprietary firm-level or trading data informs the empirical work. Forward-looking statements in this paper are expressions of academic opinion subject to significant uncertainty and should not be relied upon for investment or financial decision-making. Past performance is not indicative of future results.

Overview

- Part 1: Background
  - Why AI exposure is measured task by task, and how LLMs came to do the scoring
  - The Eloundou et al. measure (2024)
  - Why the Eloundou measure is unstable
- Part 2: A protocol for rater-invariant measures from LLMs
  - Three properties a measure should have: repeatability, reproducibility, trueness
  - A protocol to secure them: design > test > validate
- Part 3: Applying the protocol to our measure of AI task exposure, C*
  - The protocol applied step by step, then what a task-level measure adds

How AI affects the economy: it starts at the task level



- AI acts on tasks first: an occupation is a bundle of tasks (Autor, Levy and Murnane 2003)
- Exposure measures are built at the task level, using an LLM
  - A language model scores each task; the scores are averaged up to occupations and sectors
  - The LLM has become the measurement instrument: its scores enter regressions on employment, wages and adoption, and policy reports

Measuring AI impact at the task level: The Eloundou et al. measure

- The Measure: Eloundou, Manning, Mishkin and Rock (2024, Science): GPT-4 labels every O*NET task (human annotators label a subset)

"Could an LLM cut the time to do this task by at least half, compared to a human at the same quality?"

No	Yes	Yes
E0: not exposed	**E1: directly exposed**	**E2: exposed with software**
no meaningful time saving	the LLM alone halves the time	LLM plus software built on it (E3: with image capability)

Occupation exposure: β = share E1 + ½ share E2.

Headline: about 80% of US workers have at least 10% of their tasks exposed to AI

- The issue: Yin, Vu and Persico (2026) re-run the same rubric with GPT-5, Gemini 2.5 and Claude 4.5 on the same tasks
  - Share of tasks rated directly exposed (E1): 14% under Gemini 2.5, 51% under Claude 4.5 (3.6×)
  - Agreement across the three model pairs: 57–73% of tasks (κ 0.36–0.56); exposed vs not: 81–89%, so the instability is mainly E1 vs E2
  - Employment difference-in-differences (exposure × post-ChatGPT): individual-level coefficients vary 2.4× (Gemini vs the original GPT-4)
  - County level: significantly negative under one model, insignificantly positive under another

Eloundou AI impact measure has been widely used

More than 50 studies and official products since 2023, from central banks and the BLS to top journals

Central banks (6)	Statistical agencies, government, international (8)	Policy institutes (10)	Research: labor and macro (14)	Research: firms, finance, global (14)
St. Louis Fed (2025)	US BLS: official AI exposure categories (2026)	Penn Wharton Budget Model (2025)	Acemoglu (Economic Policy 2025)	Eisfeldt, Schubert & Zhang
Philadelphia Fed (2025)	BLS–Census: firm AI adoption (Dey et al. 2026)	Yale Budget Lab (2025; 2026)	Brynjolfsson, Chandar & Chen (2025)	Hui, Reshef & Zhou (Org. Science)
Dallas Fed (2026)	California Policy Lab with CA EDD: UI claims (2026)	Brookings (2024; 2025)	Humlum & Vestergaard (PNAS)	Bertomeu et al. (Italy ChatGPT ban)
Cleveland Fed (Rinz 2026)	Idaho Department of Labor (2026)	Stanford Digital Economy Lab indicators	Frank, Ahn & Moro (PNAS Nexus)	Liu et al.
Deutsche Bundesbank (Falck & Röhe 2026)	IMF: AI and productivity in Europe (2025)	Tufts Digital Planet (2026)	Labaschin et al. (AEA P&P)	Chen et al. (China Econ. Review)
Bank of England (2026, robustness)	OECD: G7 productivity projections (2025)	CEP (Lambert & Schindler)	Hosseini & Lichtinger	He et al. (China)
	OECD: Miracle or Myth? (2024)	IPPR, UK (2024)	Azar, Giné & Sanz-Espín	Mishra (India)
	IDB: Latin America (Azuara et al.)	ETLA, Finland (2023)	Hernæs & Kostøl (IZA)	Murugan et al. (global)
		Funcas, Spain (2026)	Kauhanen & Rouvinen (Finland)	Tomlinson et al. (Microsoft)
		EsadeEcPol, Spain	Henseke et al. (UK)	del Rio-Chanona et al.
			Klein Teeselink & Carey	Bagnoli, Bouquet, Kaboli & Sheffi
			Johnston & Makridis	Gupta & Kumar; Schaal; Garg et al.
			Chen (2026); Steele & Cruz	

- Their headline numbers rest on these labels: AI adds at most 0.66% to TFP over ten years (Acemoglu); +0.2 to 1.3 pp a year G7 productivity (OECD); −13% employment for 22–25-year-olds in exposed jobs (Brynjolfsson et al.)

Why might the Eloundou measure be unstable? Four potential pitfalls

- 1) It asks the model to estimate a latent quantity
  - Would AI cut this task's time by 50%? Nothing in the task statement says; each model must reconstruct it, and can confabulate
- 2) The quantity is a ratio of two estimates
  - Time without AI ÷ time with AI: errors in both compound, most near the threshold
- 3) The cut is placed without checking the distribution
  - A hard 50% threshold on a continuum: if mass sits near it, small differences flip labels
- 4) Both the instrument and the target move
  - AI capabilities change with each release: the model is both the instrument and the thing measured
- Their human check could not catch this
  - Humans rated ~2,000 work activities; those labels were mapped onto the 19,265 tasks
  - GPT-4 and these human-derived labels agree on only 65.6% of tasks, and both estimate the same latent ratio, so neither is ground truth
  - The model was never varied

Before the protocol: three properties a measure should have

Property	A ruler fails it when…	An LLM fails it when…	Checked by
1) Repeatability (within rater)	it is elastic: the same object reads differently each time	the same model and prompt give a different score on a re-run	a within-model noise floor
2) Reproducibility (inter-rater)	two rulers of the same design disagree	a different model, or a reworded prompt, gives a different score	a cross-model test on the coefficient the measure feeds
3) Trueness (criterion validity)	an inch is missing: consistent, and wrong	all models agree, but on the wrong construct	a benchmark built without any language model

- Borrowed from two traditions that arrived at the same three properties
  - Metrology: ISO 5725 defines the accuracy of any measurement method as precision (repeatability, reproducibility) plus trueness
  - Psychometrics: reliability (test-retest, inter-rater) is necessary but not sufficient for validity (Cronbach and Meehl 1955; Campbell and Fiske 1959)
- Agreement across models is reliability, not validity: models trained on overlapping data can share errors (a model monoculture)

Before the protocol: the properties as a measurement model

Score of task t by model r on run k:

$$s(t, r, k) = \theta(t) + \eta(t) + \delta(t, r) + \varepsilon(t, r, k)$$

s: the score task t gets from model r on its k-th run · θ: the construct · η: error shared by every model · δ: model-specific error · ε: run-to-run noise

Term	Property	What can detect it
ε	Repeatability	Scoring the same tasks twice with the same model
δ	Reproducibility	Varying the model and comparing the downstream coefficient
η	Trueness	Only a benchmark that no language model produced

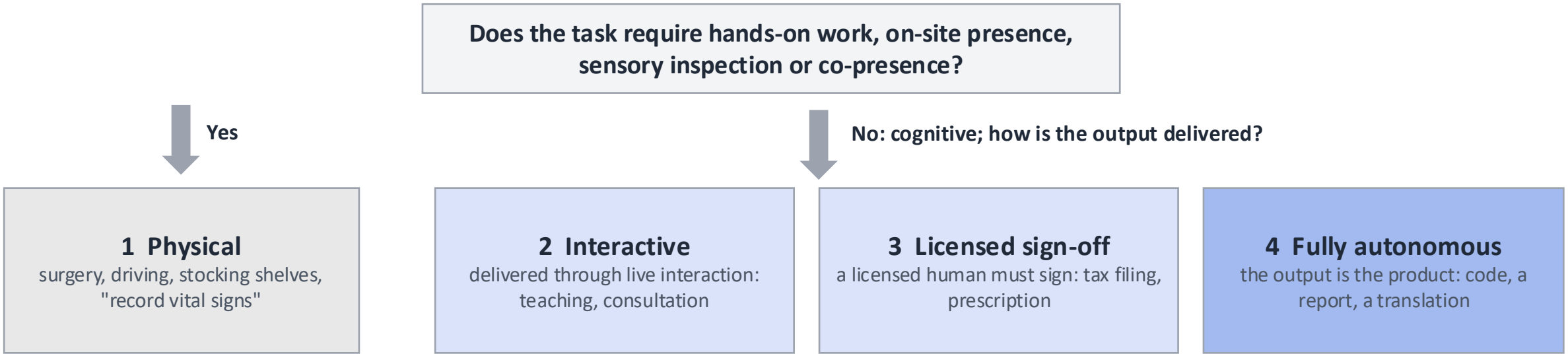
- Aggregating tasks to occupations averages out ε and part of δ
- A downstream slope $\beta(r) = \mathrm{Cov}(y, \bar{s}_r) / \mathrm{Var}(\bar{s}_r)$ can differ across models only through δ, since θ and η are the same for every model: comparing models reveals δ
- η moves every β(r) together: invisible to any comparison across models, hence a separate validation step

The protocol: three steps for three properties

Step	What you implement or test	Property it secures
1. Design	1)Choose a criterion that reads an observable property, not a latent one 2)Place the cut where cases are sparse:	Reproducibility (built in before any scoring)
2. Test	1)Test within-model noise floor 2)Test across models, and across prompt wordings 3)Test equality of the downstream coefficient the measure feeds, with inference for shared items and an equivalence margin	Repeatability (the noise floor) Reproducibility (across models and wordings)
3. Validate	Certify every model's scores against a benchmark that no language model produced	Trueness

- Agreement alone is not enough: reliability (steps 1–2) is necessary, but only step 3 can catch errors all models share

Our AI task exposure measure: C*



- Each of the 17,536 O*NET task statements is scored 1–4 by an LLM, with an identical prompt for every model
- C* measures workflow completeness: what a task needs from a human, from a body (1) to nothing, because the output is information (4)
- From task to occupation score: O*NET importance-weighted mean of its tasks; the task distribution is kept

C* and the Eloundou rubric side by side

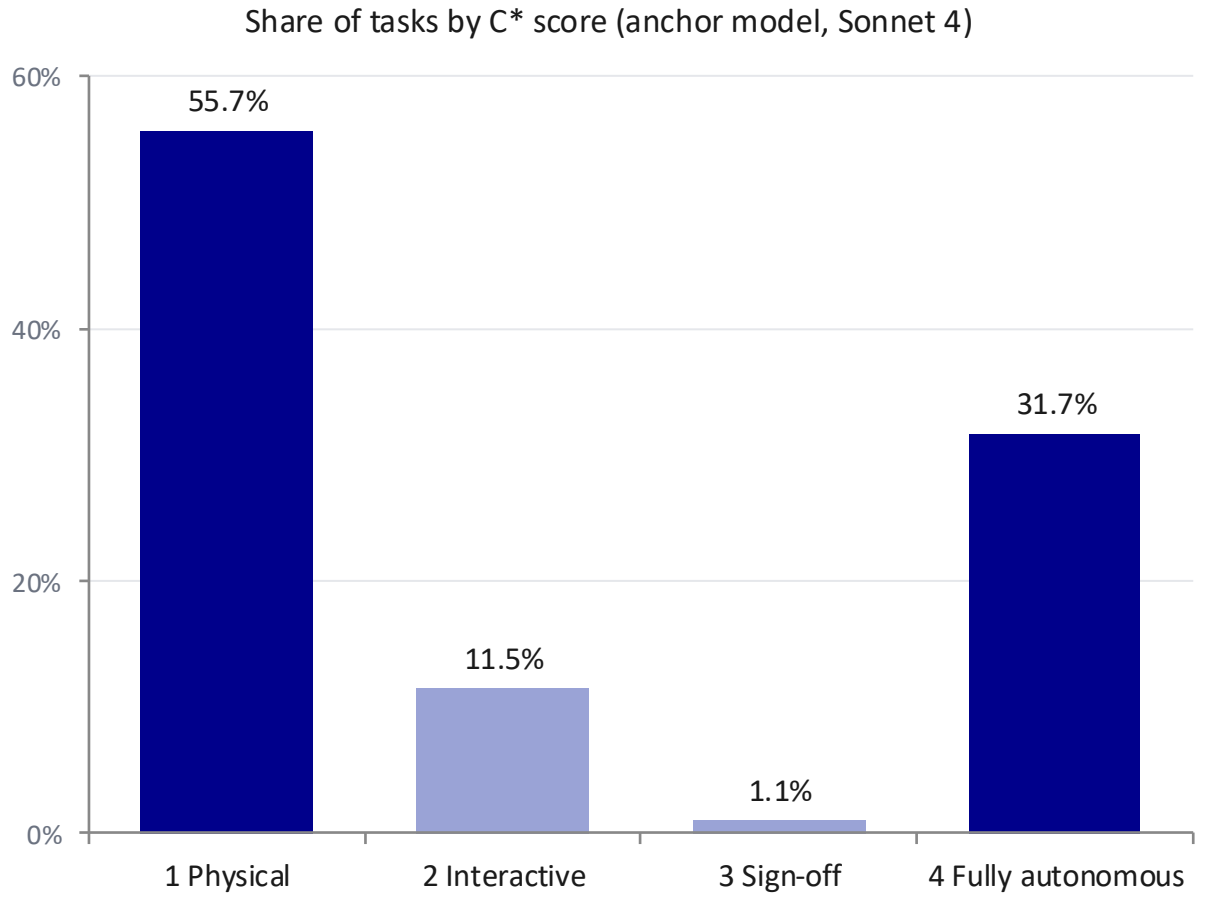
	Eloundou et al. (2024)	C* (this paper)
Question asked	Could an LLM (± software) cut the time for this task by at least half, at equal quality?	Does the task require physical presence? If not, how is the cognitive output delivered?
What the model must do	Estimate a latent ratio: time without AI vs time with AI	Read a property stated in the task
Labels	E0 not exposed · E1 direct · E2 via LLM plus software built on it (E3 with image capability)	1 physical · 2 live interaction · 3 licensed sign-off · 4 fully autonomous (the output is the deliverable)
The cut	A 50% time-saving threshold on a continuum	Physical vs cognitive; ambiguous interior buffered (2–3)
Moves with model capability?	Yes: defined against what current models can do	Designed not to: a property of the work
Occupation score	β = share E1 + ½ share E2; core tasks weighted 2, supplemental 1	Mean weighted by O*NET task importance, and the task distribution

The protocol applied to C*

- Step 1: Design
  - C*'s observable criterion, and a check on where tasks fall
- Step 2: Test
  - Test 1, repeatability: within-model noise floor
  - Test 2, reproducibility upstream: agreement across models
  - Test 3, reproducibility downstream: equal coefficients across models, with inference for shared items and an equivalence margin; one prompt-wording check
- Step 3: Validate
  - Trueness: a pre-2022 human survey with no LLM in it (AIOE, n = 683 occupations); firm-reported adoption (BTOS n=18)

Step 1, Design: read something observable

- 1) Focus on observability: whether a task needs physical handling is written in the task statement
  - "Perform surgery", "Take and record patient vital signs": every model reads the same fact
  - A property of the work, not a forecast of AI capability, so it does not move with model releases
- 2) Partitioned with care: physical gate first, then a cognitive sub-scale
  - Ambiguous tasks get their own categories (2–3) rather than being forced across the cut
  - The validated object is the binary: physical (1) vs cognitive (2–4)
- Check: 86% of tasks score 1 or 4, the two poles (82.7–88.3% across the six models)
  - So the cut falls where tasks are sparse; a consistency check, not evidence on its own



*Shares of the 17,536 unique O*NET task statements, each counted once (no occupation or employment weights).*

Step 2, Test 1: repeatability within a model

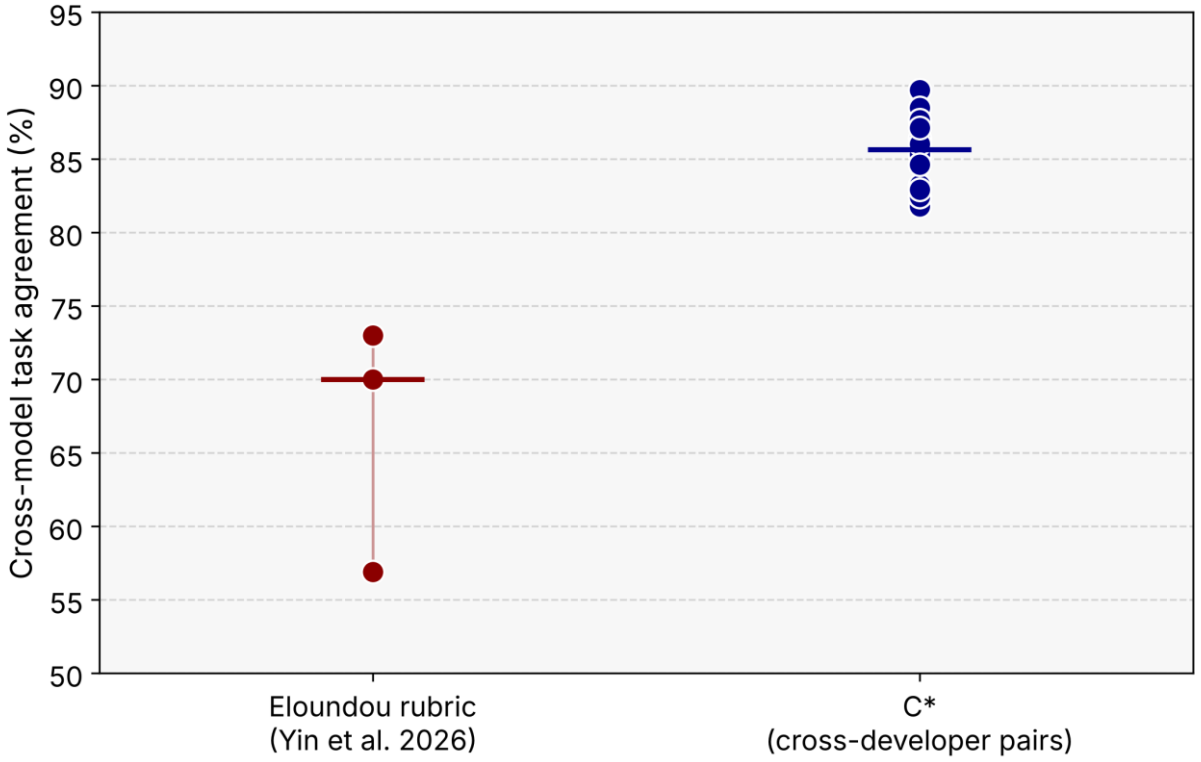
- **Testing: does each model agree with itself on a re-run? (repeatability)**
- Six models, four developers
  - Three vintages within Anthropic; then OpenAI, Google, DeepSeek
- Held fixed: byte-identical prompt, task set, aggregation pipeline
  - Temperature at provider default: the floor bounds its random noise, not systematic differences
- Noise floor: 1,000 tasks scored twice, self-agreement 90.9–98.6%

Model	Developer	Self-agreement
Sonnet 4.5	Anthropic	98.6%
Opus 4.8	Anthropic	97.4%
DeepSeek-V3	DeepSeek	96.3%
Sonnet 4	Anthropic	96.0%*
GPT-5	OpenAI	95.1%
Gemini 2.5 Pro	Google	90.9%

*Self-agreement: share of 1,000 tasks given the same score on two runs. *Cached scores vs a 982-task re-score before retirement.*

Step 2, Test 2: reproducibility across models

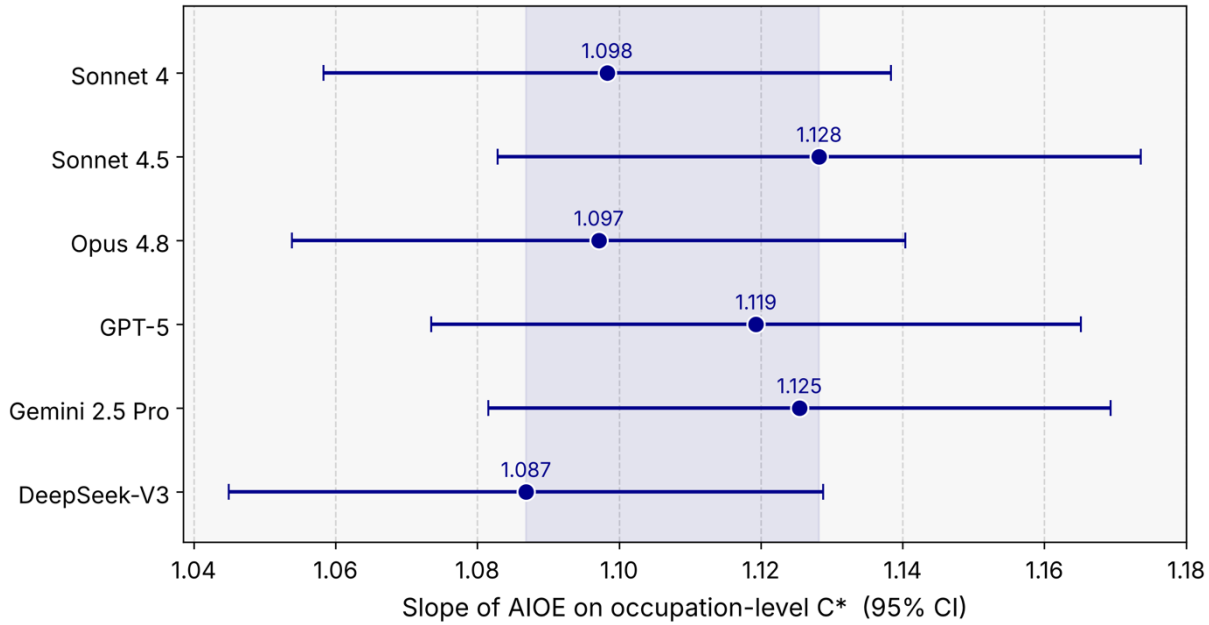
- **Testing: do different models give the same task labels? (reproducibility)**
- C*: all 12 cross-developer pairs agree on 81.8–89.7% of tasks (κ 0.68–0.82); Eloundou (Yin et al., 3 models): 57–73%, or 81–89% on exposed vs not
- Agreement is highest where the criterion is most observable:
  - Physical (vs cognitive: Fleiss κ 0.837; dropping any one model: 0.826–0.843
  - One model per developer: κ 0.838 (all six: 0.837)
  - Licensed sign-off (score 3): κ 0.33, but also rare (0.7–8.1% of tasks), so not a clean test of observability
- It averages out: occupation-level correlation across models 0.95–0.98; sector level 0.997–0.999



Each point is a model pair; ticks are medians. Eloundou values from Yin et al. (2026), Table 1.

Step 2, Test 3: reproducibility of the downstream coefficient

- **Testing: does the coefficient the measure feeds depend on the model? (reproducibility, downstream)**
- AIOE: occupational AI exposure from a pre-2022 human survey (Felten, Raj and Seamans 2021); no LLM in it. It is a pre-2022 human survey with no language model in it
- AIOE regressed on each model's occupation-level C* (n = 683): slopes 1.087–1.128
- Small: the largest pairwise gap is 3.7% of the mean slope
  - Clustered 95% CI 1.8–5.7%; across all six, the bootstrap upper bound is 6.3%
- But not zero: the test is precise enough to reject exact equality
  - Stacked regressions, occupation-clustered: $\chi^2(5) = 23.1$
- **So: statistically detectable, economically small**
- Allowing for picking the largest gap (occupation bootstrap): equivalent for any margin above 6.3%; ±10% is a convention



Slope of AIOE on occupation-level C, one regression per model; occupation-clustered 95% intervals.*

Step 3, Validate: trueness against a benchmark no LLM produced

- **Testing: does the measure track a benchmark no language model produced? (trueness)**
  - Same AIOE data as Test 3: there we compared models with each other, here each model with the benchmark
- AIOE (Felten, Raj and Seamans 2021): a pre-2022 human survey with no language model in it
  - It checks the structure C* reads, not generative-AI exposure as such
- Every model correlates 0.84–0.87 with AIOE, and none validates detectably worse
- AIOE's main axis is also physical vs cognitive, so the test that matters is the signal beyond physicality (slide 26)

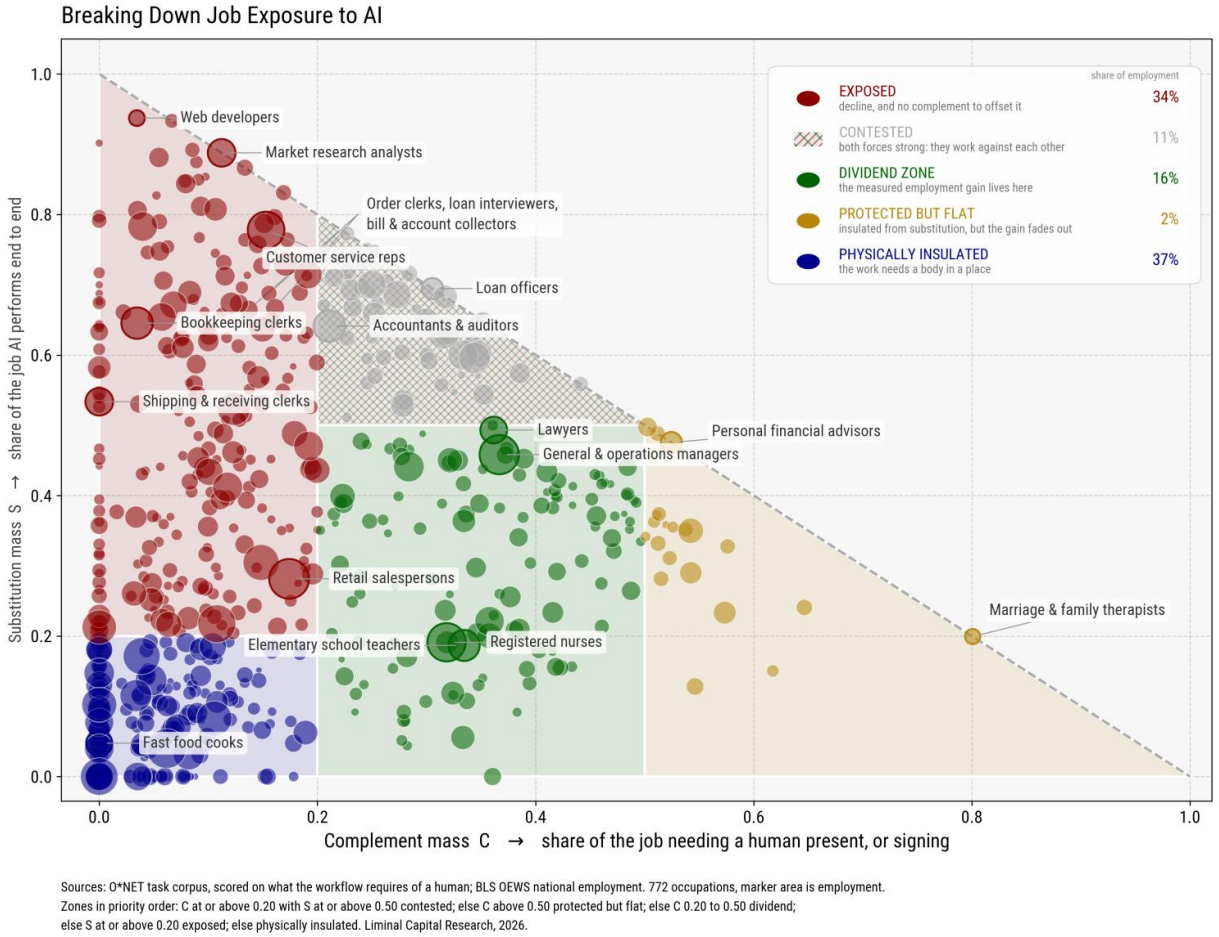
Benchmark	n	Correlation with C*
AIOE, occupations	683	0.844–0.870 across the six models
AIOE, beyond a 17-element physical-ability composite (partial)	682	0.37 [0.31, 0.43]
Firm-reported AI adoption, BTOS (NAICS-2 sectors)	19	0.80–0.82 across the six models

Protocol summary

Step	Property	Result for C*
1. Design	Reproducibility by construction	Reads an observable property; 86% of tasks at the poles
2. Test	Repeatability	Each model agrees with itself on 91–99% of tasks
	Reproducibility, upstream	Cross-developer agreement 82–90% (Eloundou, Yin et al.: 57–73%); binary κ 0.84
	Reproducibility, downstream	largest AIOE slope gap 3.7% [1.8, 5.7]; equivalent for any margin above 6.3% (Eloundou β, same six models: 41%)
3. Validate	Trueness	r = 0.84–0.87 with AIOE for every model (Eloundou β: 0.78–0.88); partial r 0.37 beyond physicality

Application: C* as a measure of AI impact on jobs

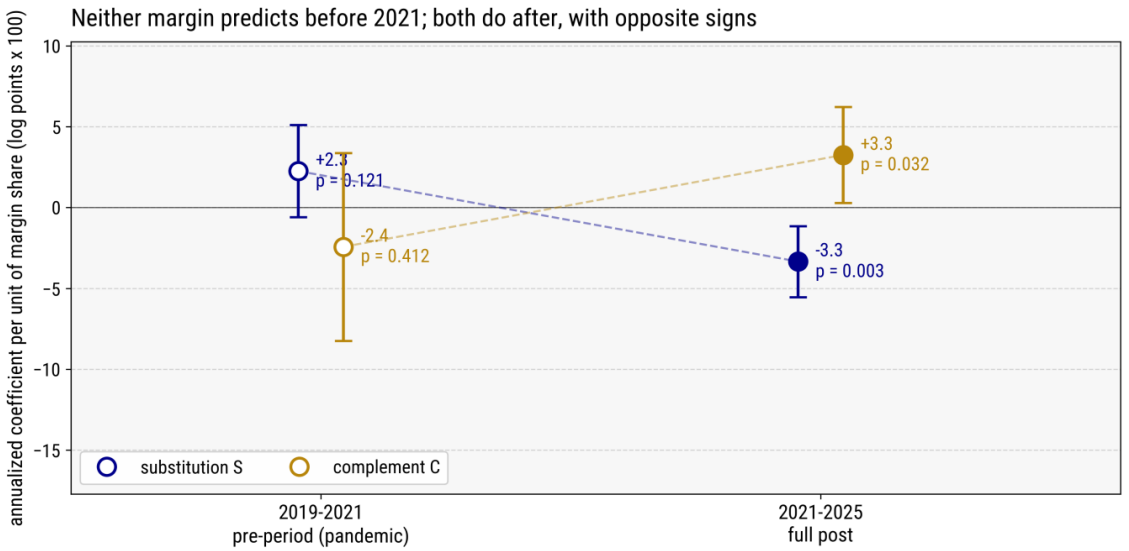
- Read C* as an economic classification of each task:
  - Score 4 → “AI substitutable”: AI performs the task end to end (y-axis)
  - Scores 2–3 → “AI-complemented”: a human present, or a licensed human signing (x-axis)
  - Score 1 → inert I: physical work AI cannot do
- Each of 798 occupations gets shares S, C and I, with S + C + I = 1
- US employment: 32% substitutable, 15% complementary, 53% physical (45M / 21M / 77M jobs)
- A single exposure number averages two margins that can push in opposite directions
- S vs C uses the split within cognitive work, less reliable than the binary (sign-off κ 0.33 sits in C)



Verschuere and Cameron (2026), Hiding in the Mean.

Extra: Using C* in a DiD

- Annualized log employment growth on S and C together, physical share omitted (705 occupations, employment-weighted, log-wage control, HC1)
- 2019–2021: neither margin predicts employment growth
- 2021–2025 levels: −0.033 on S (p = 0.003), +0.033 on C (p = 0.031)
  - Substitutable work grew more slowly; complemented work grew faster
  - This is the level, not the DiD: in the DiD, C is +0.056, p = 0.114
- C and S have opposite signs: a single exposure index nets them out



Sources: OEWS national employment 2019, 2021 and 2025; margins from the task-score cache; all estimates computed by this script and equal to Table 2's rows. Annualized log employment growth on S and C in one regression, winsorized 1/99, employment-weighted WLS with a log-wage control and HC1 errors, full panel. Filled marker means p below 0.05; bars are 95 percent intervals. The 2019-21 window is a pandemic pre-period and not a placebo: its substitution sign is positive rather than zero, and Section 7.1 prices what that does to the difference between the columns, which is the canonical DiD (-0.056 on S at p = 0.003, +0.056 on C at p = 0.114). The pre-AI 2015-19 placebo is null on both margins (Table 8). Liminal Capital Research, 2026.

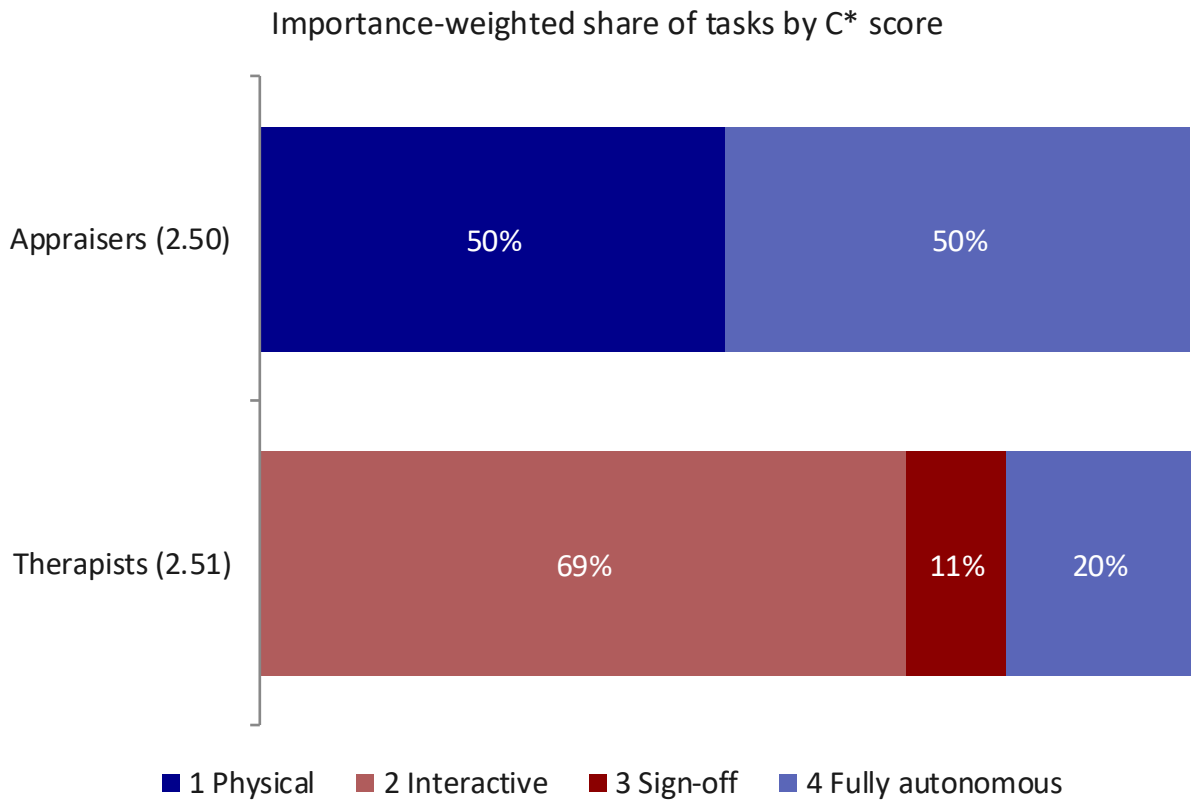
Verschuere and Cameron (2026), Hiding in the Mean.

Limitations and Open Questions

- What agreement cannot see
  - Error shared by all models (η): six models, four developers, overlapping training data; κ and the slope test condition on this panel
- Weak spots in the classes
  - Licensed sign-off (score 3) is model-dependent (κ 0.33; 0.7–8.1% of tasks across models)
- What validation covers
  - AIOE validates rankings, not levels, and shares the physical axis (partial correlation beyond physicality 0.37)
  - No human ground truth yet: a random, multi-coder gold sample is next
- Open questions
  - How should models be combined into one measure: majority, median, or soft labels?
  - What to do with the licensed sign-off class (3): keep it, or group it into one human-in-the-loop class

Discussion: why a task-level measure? The mean hides the shape

- Same mean, opposite composition:
  - Appraisers (personal and business property), mean 2.50: half physical inspection, half write-up
  - Marriage and family therapists, mean 2.51: mostly live interaction
- Under imperfect substitution the shape governs the outcome; in the Leontief limit the bottleneck task sets the ceiling
- Mean within-occupation SD of task C* is 0.98; 83% of occupations mix a physical and a fully autonomous task (score 4)
- 74.0% of workers have ≥10% of tasks fully autonomous (score 4); 26.1% have more than half



Discussion: is C* just physicality, or a keyword rule?

- Mostly physicality, by design: AI exposure is first of all cognitive vs physical work (C* vs occupational physical share: r = −0.97)
- But not only physicality: partial correlation of C* with AIOE
  - After a 17-element O*NET physical-ability composite: 0.37 (95% CI 0.31–0.43, n = 682)
  - After Dingel–Neiman teleworkability: 0.75 (n = 683)
- And not a keyword rule: a physical-verb regex
  - Agrees with the six-model consensus on 64.0% of tasks (κ 0.31), vs 92.6–95.2% for each model
  - Its partial with AIOE after the physical control: 0.05 (p = 0.21); the consensus keeps 0.38
- Observable is not trivial: models agree because they read what the task involves, not which words it contains

Conclusion

- **Treat an LLM as a measurement instrument that must be validated, not as an oracle**
- Follow a clear protocol:
  - 1) Design: have the model read something in the task, not estimate something that is not there
  - 2) Test: compare models against a within-model noise floor, on the coefficient the measure feeds, with an equivalence margin
  - 3) Validate: certify every model against a benchmark no language model produced
- C* is an existence proof: a task-level exposure measure whose physical/cognitive split and substitution margin survive the choice of model; each new model release tests it again
- Paper: SSRN 7114299 · ben.verschuere@liminal-capital.com