RISEI LAB · NORTHWESTERN UNIVERSITY

POLICY BRIEF · NO. 2026-03

The Adaptive Precision Framework: From Static Measurement to Continuous Recalibration in the AI Era

Every major forecast about which jobs AI will eliminate comes from asking a large language model to rate itself. We found the answer depends entirely on which model you ask. If the measurement is unstable, the policies built on it are too. This brief presents what we found and what it means for workforce and education policy.

Michelle Yin, Ph.D.
Associate Professor, School of Education and Social Policy
Founding Director, RISEI Lab
Northwestern University

April 2026

This brief summarizes findings from "How (un)Stable Are LLM Occupational Exposure Scores? Evidence from Multi-Model Replication" (Yin, Vu, & Persico, 2026, RISEI Working Paper Series No. 4152026).

WHAT WE FOUND

- **The scores are not stable.** Researchers measure AI's impact on jobs by asking a large language model to rate which occupational tasks it could speed up. We replicated this approach with three different models on the same tasks. The results varied 3.6-fold. The lowest agreement between any two models was 57%.
- **The instability changes conclusions.** When we used these scores to estimate employment effects, the results depended on which AI model generated them. One model indicated job losses in a county. Another indicated no effect. Same data, same method, opposite conclusions.
- **A feedback loop makes it worse.** Occupations where more people already use AI showed bigger increases in measured "exposure" over time. The scores are shaped by who uses the technology, not just what the technology can do.
- **Major institutions rely on these scores.** The Bureau of Labor Statistics, OECD, International Labour Organization, and World Economic Forum all use occupational exposure scores to project employment, allocate retraining funds, and design workforce policy. If the scores shift 3.6-fold, those decisions inherit that fragility.

THE MEASUREMENT PROBLEM AT A GLANCE

3.6×

Ask a different AI, get a 3.6× different answer about which jobs are at risk

Sign Flip

One AI says jobs are disappearing. Another says they're fine. Same data.

<5 yrs

Half of what you learned 5 years ago is already outdated. In tech: 2.5 years.

16%

We're making policy for the world based on data from 1 in 6 people

Source: Yin, Vu, & Persico (2026, RISEI Working Paper No. 4152026); WEF (2025); Microsoft (2026).

A rapidly growing body of research uses AI-generated scores to predict which jobs will be most affected by artificial intelligence. Our research shows these scores are highly unstable. This brief explains how the scores work, what we found, why it matters, and what workforce and education systems should do about it.

How Occupational Exposure Scores Work

The U.S. government's O*NET database describes every occupation as a list of tasks. A nurse, for example, has tasks like "monitor patient vital signs" and "document care plans." There are roughly 19,000 such task descriptions across about 800 occupations.

Researchers ask a large language model (LLM), such as ChatGPT, to read each task and judge whether it could help a worker complete that task at least twice as fast. Each task gets one of three ratings:

THE RATING SYSTEM (ELOUNDOU ET AL., 2024)

- **E0** **Not exposed.** AI cannot meaningfully speed up this task. Example: "Physically restrain a combative patient."
- **E1** **Directly exposed.** An LLM alone could cut the time in half. Example: "Draft a patient discharge summary."
- **E2** **Exposed with tools.** An LLM plus other software could cut the time in half. Example: "Analyze trends in patient lab results."

The share of an occupation's tasks rated E1 becomes its "occupational exposure score." This number has become the standard way researchers and policymakers measure how much AI might affect a given job. The original paper (Eloundou et al., 2024), published in *Science*, has been cited in over 1,500 studies. The scores are used directly by the Bureau of Labor Statistics in its 10-year employment projections (Machovec, Rieley, and Rolen, 2025), by the OECD to assess AI's impact across member countries (Green, 2024), by the International Labour Organization to estimate global effects (Gmyrek, Berg, and Bescond, 2023), and by the World Economic Forum in its Future of Jobs Report (WEF, 2025).

The problem: the LLM is rating its own capabilities. If a different model is asked the same questions, it gives different answers.

What We Found

We applied the same rating rubric to the same 18,797 tasks using three different large language models released in 2026: ChatGPT-5, Gemini 2.5, and Claude 4.5. The results diverged sharply.

3.6×	57%	2.4×	±
How much occupational exposure scores diverge across large language models	Worst-case agreement between any two AIs rating identical tasks	How much estimated job-loss effects vary by model choice	County-level results flip from "job loss" to "no effect" by model

Source: Yin, M., Vu, H., & Persico, C. (2026, RISEI Working Paper No. 4152026). Northwestern University.

One model (Gemini) rated only 14% of tasks as directly exposed to AI. Another (Claude) rated 51%. Their agreement on individual task ratings was poor. When we used each model's scores to estimate whether AI exposure is associated with employment changes, the choice of model determined the conclusion. At the individual worker level, the estimated effect ranged 2.4-fold. At the county level, one model showed a statistically significant association between AI exposure and job losses. The original scores and two other models showed no significant effect. Whether AI appears to be displacing jobs in a county depended entirely on which AI model generated the occupational scores.

We confirmed this disagreement is not random. When we asked the same model to rate the same tasks multiple times, it agreed with itself over 90% of the time. The disagreement is between models, not within them. Different AI systems have systematically different thresholds for what counts as "direct" AI assistance.

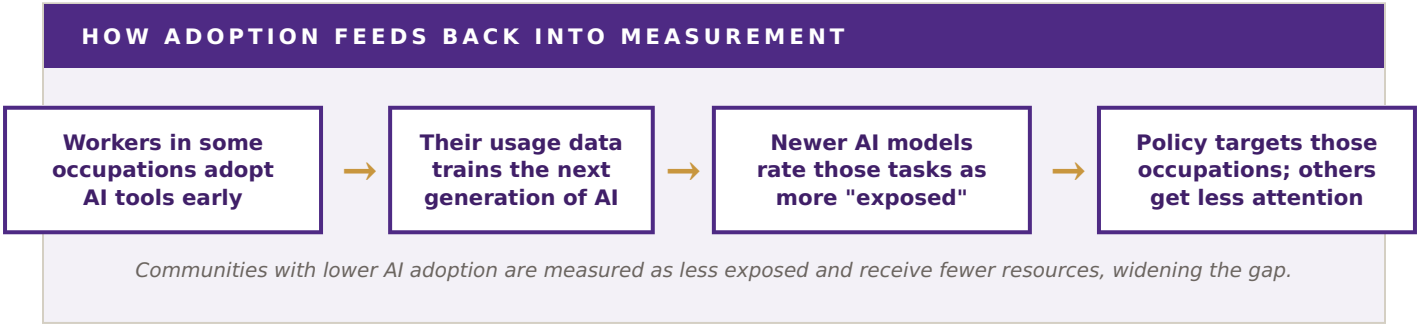
Why This Is More Than a Technical Problem

In most research, measurement error makes it harder to find effects: it pushes results toward zero. The instability we document is different. Different models do not just add noise around a true answer. They produce systematically different rankings of which occupations are most affected. The entire profile of "AI-exposed jobs" looks different depending on which AI model draws it.

This means the error in downstream studies is unpredictable. A researcher using one model may find significant job displacement. A researcher using a different model, with identical data and identical methods, may find none. Massenkoff and McCrory (2026) independently confirmed that these capability-based scores alone do not predict actual employment trends in Bureau of Labor Statistics data. Only when combined with data on how workers actually use AI tools does the measure gain predictive power.

The Feedback Loop

The measurement problem compounds over time. We tested whether occupations where more people already use AI tools also show larger increases in measured exposure when newer AI models are released. They do (statistically significant at $p < 0.05$). The reason: AI models improve by learning from how people use them. When more people in a given occupation use AI, the next generation of models gets better at those tasks, and rates those tasks as more "exposed." Policy attention then flows to those occupations, while communities with lower adoption are measured as less exposed and receive less support.



WHOSE WORK IS REFLECTED IN THESE SCORES?

The occupational exposure scores reflect the capabilities of AI models trained primarily on the work patterns of English-speaking, urban, college-educated populations. Only **16.3% of the world's population** has ever used a generative AI tool (Microsoft AI Economy Institute, 2026). In the U.S., daily AI use is under 20% of adults (Bick, Blandin, and Deming, 2025). Global workforce policy is being built on scores calibrated to a small segment of the workforce.

What This Means for Policy

If the occupational exposure scores are unstable, every policy decision built on them is fragile. Training programs targeting "high-exposure" occupations are targeting a list that changes depending on which AI model generates it. Employment projections incorporating these scores inherit their instability. The most basic policy question, how many jobs will AI affect, cannot currently be answered with confidence. Our county-level results show that the answer is not just uncertain in size; it can reverse direction entirely.

What Should Be Done

1 Require Multi-Model Testing

No study or policy should rely on scores from a single AI model. Results should be reported across at least three models. The range of disagreement should be treated as a measure of genuine uncertainty.

2 Invest in Usage-Based Measures

Scores based on what AI *can* do are not the same as what workers actually *use* it to do (Massenkoff & McCrory, 2026). The field needs measures grounded in real-world adoption, not hypothetical capability.

3 Audit the Feedback Loop

Track whether the scores systematically overrate occupations with early AI adoption and underrate those with lower adoption. Flag when measurement bias amplifies existing inequalities.

4 Treat Scores as Time-Stamped

AI capabilities change with every model release. Any analysis using these scores should state which AI model was used, when, and acknowledge that a different model would produce different results. Static scores should not drive multi-year policy.

Implications: A Perspective

The following reflects the author's interpretation of these findings and their implications for workforce and education policy. These are policy conclusions drawn from the evidence, not findings of the paper itself.

This paper documents instability in the scores used to predict which jobs AI will affect. But the relevant question for practitioners is not whether the scores are stable. It is what employers and educators should do in a world where they are not.

FOR THE WORKFORCE

The empirical evidence on AI's actual effects on jobs reinforces the measurement concern. Despite rapid adoption, several rigorous studies find little aggregate employment impact so far. The Budget Lab at Yale finds no clear relationship between occupational AI exposure and unemployment through August 2025 (Gimbel et al., 2025). Humlum and Vestergaard (2025), using Danish government records, find essentially zero effects on earnings or hours for workers in 11 AI-exposed occupations. Hartley et al. (2026) find that 35.9% of U.S. workers used generative AI by December 2025, with small positive wage effects but no decline in employment. The Atlanta Fed (2026), surveying corporate executives, finds that firms reported negligible AI-driven changes in headcounts in 2025, even as 60% invested in AI and reported productivity gains.

Beneath these aggregate numbers, there are signs of change at the margins. Brynjolfsson, Chandar, and Chen (2025), using payroll records covering millions of workers, find reduced hiring of entry-level workers in AI-exposed occupations. Firms appear to be adjusting by restructuring tasks and shifting workforce composition rather than making broad layoffs. Stephany et al. (2026) find that AI skills on a resume help offset age and education disadvantages in hiring decisions.

A critical caveat: nearly all of these studies define "AI-exposed" occupations using scores from a single AI model. Our paper shows these scores vary 3.6-fold across models, and the estimated employment effect can reverse sign. The null findings therefore admit two readings: AI is not yet displacing workers in aggregate, or the occupational scores used to identify affected workers are too imprecise to detect displacement even if it is occurring. These readings have very different policy implications. Until studies routinely test their results across multiple AI models, the workforce evidence should be read with this uncertainty in mind.

Given this ambiguity, workforce policy should focus on:

Monitoring changes at the margins, not just averages. Overall employment levels may remain stable while entry-level hiring, career ladders, and task assignments shift substantially. Workforce agencies should track hiring rates, task reallocation, and career entry by occupation and demographic group.

Moving from static job descriptions to ongoing assessment. Most employers still write a job description once and use it for years. They evaluate employees annually against criteria set months earlier. Every one of these practices assumes the skills needed today will be the skills needed next quarter. When occupational profiles can shift with every AI model generation, that assumption is difficult to defend.

Investing in transition support, not prediction. If we cannot reliably predict which occupations will be most affected, policy should invest in infrastructure that helps workers through transitions regardless: portable benefits, rapid retraining, income support during skill building, and credentialing tied to demonstrated ability.

Addressing the skills verification gap. There is a parallel instability at the individual level. The number of unique credentials in the U.S. has grown from 334,000 in 2018 to over 1.85 million in 2025 (SHRM, 2026), making it harder for employers to distinguish which credentials signal real competence. Eighty-five percent of companies say they practice skills-based hiring, but 53% lack standardized processes to support it (Fortune, 2026). Education credentials are losing signaling value at the moment when reliable signals matter most. The field needs performance-based assessments that validate what a candidate can do, not just what programs they completed.

FOR EDUCATION

The evidence gap in education is, if anything, wider than the measurement gap in labor economics. Stanford's SCALE Initiative reviewed over 800 studies on AI in K-12 education and found that just 20 met the standard for rigorous causal evidence (Stanford SCALE, 2026). None of those 20 were high-quality studies of student AI use in U.S. K-12 classrooms. Among the studies that do exist, results are mixed: student performance sometimes improves while using AI tools, but in several cases it is unchanged or declines once the tools are removed. There is almost no research on long-term learning, on equity and access, or on how AI affects different groups of students.

Meanwhile, adoption has moved far ahead of the evidence. Sixty-three percent of K-12 teachers have incorporated generative AI into their work, a 12-point increase in one year (Cengage Group, 2025). Fifty-four percent of students use AI for school (Doss et al., 2025). Eighty-eight percent of students use generative AI for assessments. Yet 68% of teachers received no formal training, and over half of school districts have not aligned AI with their mission or strategic plan (Michigan Virtual, 2025). In practical terms, this is the largest uncontrolled experiment in education history. The technology is extraordinary. The evidence base is thin. The stakes are children.

Our measurement findings connect directly. If the occupational scores used to predict which skills AI will change vary 3.6-fold across models, and if the evidence on whether AI tools actually improve learning is inconclusive even under controlled conditions, then education systems are making two bets on two uncertain foundations at once: that we know which skills AI will change, and that AI tools will help students learn those skills. Neither bet is currently supported by strong evidence.

This argues for a clear sequencing. Before scaling AI tools whose effects are unproven, institutions should invest in the foundational capacities that the Stanford review identifies as essential: the ability to evaluate information critically, to learn independently, and to exercise judgment when the evidence is ambiguous. Build the muscle first. Then hand them the tool. Specific priorities include:

Causal research before large-scale adoption. Zero high-quality causal studies of student AI use exist for U.S. K-12 classrooms. Eighty-eight percent of students already use AI for assessments. This gap should be treated as an emergency. Randomized trials of specific tools, in specific contexts, with outcomes measured both during and after AI use, should be the immediate research priority.

Focus on what AI removes, not just what it adds. When student performance improves during AI-assisted tasks but falls after the tools are taken away, the tool may be completing the task without building the underlying skill. A writing assignment is designed to develop the ability to organize thought. If AI drafts the essay, the thinking never happened. The question is not whether AI produces a better essay but whether students who use AI become better thinkers.

Continuous assessment, not periodic testing. Midterm and final exams catch learning gaps weeks or months after they form. The technology now enables real-time, ongoing assessment of what students know and whether the material being taught remains aligned with what the world requires. Institutions should build this capacity rather than relying on summative exams that arrive too late to change course.

Differentiation by age. A high school student learning to code and a first-grader learning to read face fundamentally different situations. The adoption gradient reflects this: 69% of high school teachers use generative AI compared with 33% of pre-K teachers (EdWeek, 2025). For young children, the priority

should be developing the human capabilities that all later learning depends on: reading comprehension, spoken language, social skills, and the capacity for sustained attention. These are built through interaction with people, not screens.

Train teachers in the science of learning, not tool operation. Sixty-eight percent of teachers have had no formal AI training. The instinct is to close this gap with workshops on AI tools. The evidence points in a different direction. If AI handles the mechanical parts of teaching, such as generating materials and drafting assessments, then the distinctive value of teachers is what AI cannot do: reading a room, diagnosing understanding in real time, building the trust that makes students willing to try hard things. Teacher preparation should deepen expertise in how people learn, with AI proficiency as a supplement.

Institutions that fail to build these capacities will continue to adopt tools faster than they can evaluate them.

THE QUESTIONS WE ARE NOT ASKING

The workforce and education challenges above share a common assumption: that the problem is one of prediction and adaptation. Which jobs will change? How do we retrain? How do we update curricula? But there may be a more fundamental set of questions that current policy debates are not addressing.

Should employers play a larger role in education, especially for entry-level workers? AI will change some portion of the tasks in every job. The workers most affected will be those at the beginning of their careers, in roles traditionally built around routine execution. If those roles shrink, the pipeline through which young people enter industries, learn organizational culture, and develop professional judgment narrows. Employers are not passive recipients of the workforce that education systems produce. They have a direct stake in how the next generation is prepared, and they may need to become active partners in designing what entry-level preparation looks like when routine tasks are increasingly handled by machines.

What happens to an organization when it stops hiring young people? Entry-level hiring is not just a labor market transaction. It is how institutions renew themselves. When firms reduce entry-level headcount because AI can perform the tasks those positions were built around, they do not simply become more efficient. They change their culture. They lose the energy, the questioning, and the fresh perspective that new entrants bring. They become older, more insular, more set in their assumptions. The long-run cost of that cultural shift may exceed the short-run savings from automation.

Are we defining entry-level work too narrowly? The traditional model treats entry-level positions as a way to get routine work done cheaply while giving new workers a foothold. AI challenges the first half of that equation: if a machine can handle the routine tasks, the economic rationale for the position weakens. But what if the real value of early-career workers is not their willingness to do routine work, but their ability to see what longer-tenured employees no longer see? New hires notice inefficiencies that veterans have learned to ignore. They ask questions that feel naive but expose assumptions. They bring perspectives shaped by different experiences, different technologies, different expectations. These are precisely the qualities that organizations need more of as AI takes over the predictable parts of work. Instead of asking how to preserve entry-level jobs designed around routine tasks, perhaps we should ask how to redesign entry-level roles around the things new workers do better than anyone else: challenge assumptions, import new ideas, and see the organization with fresh eyes.

These are questions, not answers. But they suggest that the workforce challenge ahead is not simply a retraining problem or a measurement problem. It is a design problem: how do we structure work, education, and the relationship between employers and the pipeline of new talent in a world where AI handles an expanding share of routine execution? The organizations and policy systems that engage with this question now will be better positioned than those still trying to predict which jobs survive.

TOWARD ADAPTIVE PRECISION

We propose the term **Adaptive Precision** for a different approach: using real-time data to continuously recalibrate what institutions teach, how they hire, how they assess, and how they design work. For employers, this means tailoring job design to match employee strengths rather than forcing people into static roles. For educators, this means making curricular and assessment decisions based on current evidence rather than assumptions locked in during the last review cycle. For both, it means treating every job description, syllabus, rubric, and hiring criterion as a living document.

None of this follows mechanically from a measurement error paper. But the measurement error documented here is a symptom of a deeper reality: AI capabilities are changing faster than the institutional systems designed to respond to them. If the ruler keeps changing, the institutions that depend on it need the capacity to recalibrate, anchored in the human capabilities that hold their value regardless of which AI model is dominant. Not a moving target, but a fixed star.

Citation: Yin, M. (2026). *The Adaptive Precision Framework: From static measurement to continuous recalibration in the AI era* (Policy Brief No. 2026-03). RISEI Lab, Northwestern University. https://sites.northwestern.edu/risei
Contact: michelle.yin@northwestern.edu · sites.northwestern.edu/risei

RISEI Lab
Research and Innovation for Social and Economic Inclusion
School of Education and Social Policy · Northwestern University · 2120 Campus Drive, Evanston, IL 60208