The AIFE Engine: A Computational Framework to Quantify and Forecast the Great Labor Reallocation

Authors: Liqian Yan, Yintong Chen, Zichun Qiu

Province: Shanghai

Country: People's Republic of China

Instructor: Mingmao Deng

Regeneron International Science and Engineering Fair

2026

Date of Completion: February 26, 2026

The AIFE Engine: A Computational Framework to Quantify and Forecast the Great Labor Reallocation

Abstract

Artificial intelligence is transforming labor markets around the world, yet we do not completely understand its effects on individuals. Previous research has used aggregated data and proxies, which fails to isolate the impact of AI on different tiers of the workforce. Thus, we investigated whether AI penetration in firms has a causal effect on the displacement of routine labor and the complementarity of creative and complex work.

Our dataset includes over 100 million different job postings from 2015 to 2024. Using this dataset, we developed a knowledge distillation system that trains an efficient student model using accurate LLM outputs as the teacher. The results were validated against 9,500 human labels. Each job posting was then categorized as routine, complex or creative and given an intensity score, producing a massive panel across 72 countries.

Our results show that one unit increase in AI intensity reduced routine labor demand by 2.1 percentage points and increased creative labor demand by 1.8 percentage points. The effects strengthened over three years after adoption and were twice as large in countries with higher income. To prove the causality of our preliminary results, we used econometric methods such as event study and instrumental variable. We also created a forecasting system using semantic embedding features and a temporal deep learning model. Compared to established baseline models, the model reduced prediction errors by 18-31%. We conclude that AI adoption is driving a structural reallocation of labor demand, as it substitutes for routine labor and complements complex and creative labor.

Keywords: Large Language Models (LLMs); Computational Social Science; Labor Market Dynamics; Time-Series Forecasting; Extreme Gradient Boosting (XGBoost); Natural Language Processing (NLP).

Contents

1 Introduction

In recent years, artificial intelligence (AI), represented by machine learning, large language models (LLMs), and automation technologies, has penetrated economic and social sectors at an unprecedented pace, reshaping the global labor market landscape. Yet, this process is inherently double-edged: AI is not only substituting repetitive labor but also spawning new occupations such as "prompt engineers" and "AI ethicists". This duality has ignited a heated and ongoing debate between policy makers, industry leaders, and scholars. Despite considerable discussion, our understanding of the current impact of AI on the labor market remains surprisingly unclear, often fluctuating between optimistic predictions and dystopian anxieties. Furthermore, the current market is characterized by a high degree of uncertainty and a high sense of urgency and requires immediate insight into the unprecedented influence of AI on the labor market.

Significant gaps persist in existing research regarding the microlevel mechanisms through which AI specifically affects labor demand. On one hand, much of the literature focuses on macro-level effects (e.g., industry-level employment substitution rates) or single mechanisms (e.g., "substitution effect" or "creation effect"). On the other hand, AI application is often measured using superficial methods like keyword searches (e.g., "AI" or "machine learning"), which fail to capture the actual depth of AI penetration in enterprises. Coupled with unresolved endogeneity issues (e.g., high-skill enterprises may proactively adopt AI and adjust labor structures), these limitations undermine the reliability of causal inferences. Additionally, critical issues such as the generative logic of emerging occupations and the dynamic changes in skill premiums still lack empirical support from microlevel data. Current research also lacks a framework that effectively predicts future AIFE trajectories.

To address these gaps, this study rigorously analyzes the realistic impact of AI on the labor market. Rather than relying on traditional proxies, we employ a large-scale dataset of over 100 million online job postings gathered from major international recruitment websites between 2015 and 2024. To accurately measure AI's influence, our paper constructs a novel indicator, "Employment AI Intensity (AIFE)", which systematically quantifies the extent to which specific occupations are susceptible to transformation driven by AI integration. Subsequently, due to the unstructured and high-dimensional nature of web-scraped data, we employ a novel, three-stage Knowledge Distillation pipeline consisting of entity resolution, teacher–student distillation with human-in-the-loop calibration, and systematic evaluation against established baseline methods. This guarantees the accuracy and applicability of our dataset. We then identify the descriptive evidence that emerges from deploying the validated AIFE pipeline on the corpus.

We further apply econometric methods to establish causality and alleviate endogeneity, thus ensuring the validity of our findings. We first construct a Two-Way Fixed Effects (TWFE) model to isolate the effects of changes in AIFE on labor composition, controlling for time-invariant firm traits and time-varying industry-level stimuli. We proceed to verify the Parallel Trends Hypothesis through implementing a dynamic event-study design, addressing the influence of heterogeneous treatment effects across cohorts. Finally, we alleviate remaining endogeneity concerns through employing a novel, Bartik-Style Shift-Share Instrumental Variable.

In response to the lack of future-oriented, predictive frameworks of AIFE trajectories, we develop a predictive pipeline for firm-level AI adoption by introducing Semantic Leading Indicators (SLIs), high-dimensional features derived from model embeddings, to detect adoption momentum that remains latent in traditional scalar scores. These indicators are integrated into a Temporal Fusion Transformer (TFT), which utilizes a novel causal-consistency regularizer to align predictions with empirical economic constraints. By pairing these forecasts with conformal prediction for distribution-free uncertainty quantification, the pipeline generates displacement early warnings that map predicted AI trajectories to specific, calibrated declines in routine labor demand.

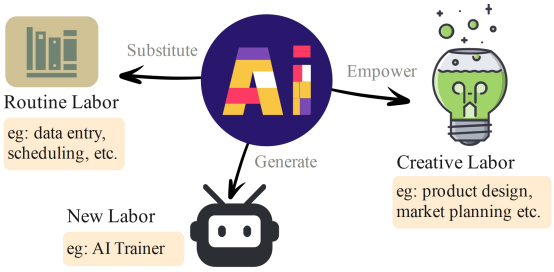

Figure 1. The Complexity of Artificial Intelligence Technologies

Our findings make clear that artificial intelligence is reshaping work. AI readily takes over routine manual and cognitive tasks, yet it also amplifies the importance of roles that require analysis, creativity, and judgment. Perhaps most striking, it is spawning entirely new kinds of jobs, both in AI-focused industries and across the broader economy. The pattern is already visible. Routine- level positions such as administrative clerks and assembly-line operators are in decline. On the other hand, demand is surging

for jobs that require higher levels of creativity, such as data analysts, research and development specialists, and AI practitioners. As this shift unfolds, workers who bring skills that complement AI are beginning to command a premium in the labor market, signaling a broader revaluation of expertise. To speak of "AI replacing jobs" is fundamentally misinterpreting and oversimplifying the situation. AI is less a job-destroyer than a job reconfigurer, transforming occupations, creating new pathways for work, and opening space for richer forms of collaboration between humans and machines.

This paper makes three contributions, structured as a progression from measurement to causal identification to prospective application.

1. **A scalable knowledge distillation pipeline for economic measurement** (§4). Our primary contribution is methodological. The AIFE pipeline distills the reasoning capabilities of a frontier LLM into a high-throughput encoder, processing 100 million documents at a total cost of $410. Systematic evaluation against eight baselines—including keyword dictionaries, published exposure indices (Webb, 2020; Felten et al., 2021), GPT-4o zero-shot, and fine-tuned BERT—demonstrates that the full pipeline achieves $\rho = 0.931$ with human consensus, a 91% improvement over keyword methods and a 16% improvement over GPT-4o zero-shot.

2. **Credible causal evidence at the firm level** (§5). We provide the first firm-level causal estimates of AI adoption on hiring composition using a global dataset. The combination of event-study dynamics (with parallel-trends validation), heterogeneity-robust difference-in-differences, and a shift-share IV strategy constitutes a multi-pronged identification approach that addresses the key threats to validity.

3. **A predictive early warning system with distribution-free uncertainty quantification** (§6). Because the causal analysis quantifies the downstream consequences of AIFE increases, predicting AIFE trajectories is equivalent to predicting labor market disruption with known effect sizes. We introduce *Semantic Leading Indicators* (SLIs), which are features extracted from the AIFE pipeline's intermediate representations that capture adoption momentum invisible to scalar AIFE histories. We embed these in a Temporal Fusion Transformer with a novel causal-consistency regularizer. Conformal prediction provides distribution-free coverage guarantees on the resulting displacement forecasts, producing actionable early warnings of the form: "Firm i will experience a Δ-unit AIFE increase (90% CI: $[\cdot, \cdot]$), implying a $\hat{\beta} \cdot \Delta$ decline in routine hiring."

The remainder of this paper is organized as follows: Section 2 reviews the literature, Section 3 and 4 detail our data and measurement systems, while Sections 5 and 6 present our causal and predictive results.

2 Literature Review

2.1 Technological Progress and the Labor Market

Early studies on the impact of technological progress on the labor market were primarily grounded in the Skill-Biased Technological Change (SBTC) framework (Katz and Murphy, 1992; Acemoglu and Autor, 2011). This theory posits that technological change tends to complement high-skilled labor while substituting for low-skilled labor, thereby leading to rising skill premiums and increased income inequality (Goldin and Katz, 2008). However, the SBTC theory struggles to explain the "job polarization" phenomenon observed in European and U.S. countries since the 1990s—where employment growth has occurred in high-skill and low-skill jobs, while medium-skill jobs have shrunk (Goos and Manning, 2007; Autor and Dorn, 2013).

To account for this, Routine-Biased Technological Change (RBTC) theory has gradually become the mainstream analytical framework (Autor et al., 2003; Acemoglu and Autor, 2011). RBTC shifts the unit of analysis from "skills" to "tasks", emphasizing that technology substitutes for "routine tasks" rather than labor with specific skill levels. Autor et al. (2003) further categorize work tasks into five types: non-routine analytical, non-routine interactive, non-routine manual, routine cognitive, and routine manual. This classification better captures the heterogeneous effects of technology on labor demand.

2.2 The Impact of Artificial Intelligence on Labor Demand

Research on the mechanisms through which AI affects labor demand focuses on several dimensions. First, the *displacement effect*: AI reduces demand for routine labor by automating repetitive, rule-based tasks. Technologies based on machine learning and natural language processing—such as OCR, RPA, and chatbots—have already been widely applied in fields like document processing, data entry, and customer service (Frey and Osborne, 2017). Graetz and Michaels (2018) found that industrial robot adoption led to significant reductions in routine manual jobs in manufacturing. In the Chinese context, Yan et al. (2020) and Wang and Dong (2020) provided similar evidence. It has been observed that AI is able to significantly replace labor demands that can be automated.

Second, the *productivity effect*: AI enhances enterprise

productivity and reduces production costs, potentially boosting labor demand (especially for non-routine cognitive jobs like R&D, management, and design) through scale expansion (Brynjolfsson and McAfee, 2014). Bessen (2019) noted that while automation substitutes for some labor, it also indirectly creates new jobs by improving total factor productivity (TFP). Different from the first effect, where AI replaces repetitive work, AI instead help to boost, but not replace, jobs that cannot be easily automated and requires cognitive thinking.

Third, the *new task creation effect*: Acemoglu and Restrepo (2018) proposed an "automation–new task balance" model, emphasizing that technology not only substitutes old tasks but also creates new ones. Generative AI (e.g., GPT, diffusion models) is giving rise to entirely new occupational fields, such as prompt engineers, AI ethicists, and data strategists (Eloundou et al., 2023). Felten et al. (2023) found that high-exposure occupations also contain substantial augmentation (rather than substitution) opportunities. Different from previous effects, which only focus on changes to current job demands, we also account for AI's effect on creating new jobs that have never existed before.

Some scholars have also highlighted the *skill restructuring and skill premium change effect*: AI application alters the structure of skill demand, driving "skill rebalancing." On one hand, demand for hard skills (e.g., programming, data analysis, AI operations) rises; on the other hand, the value of soft skills (e.g., communication, creativity, ethical judgment) becomes increasingly prominent (Deming, 2017). The emergence of generative AI may even weaken traditional "experience premiums," enabling low-experience workers to boost productivity through AI tools (Noy and Zhang, 2023).

2.3 Generative AI and Its Impact on Labor Demand

Generative AI (e.g., large language models), differs significantly from traditional AI in its impact on the labor market due to breakthroughs in natural language understanding, content generation, and complex reasoning. Generative AI doesn't just help to reduce mindpower, it also provides a direct substitution for human labor.Zeng et al. (2025) argue that generative AI may improve income distribution across groups—and even narrow income gaps—through a "capital–skill complementarity" effect.

Zhang and Dan (2025), using Chinese recruitment data, found that hiring demand for high-exposure occupations (e.g., accounting, editing, programmers) declined, while demand for low-exposure occupations (e.g., catering services, nursing) rose. Eloundou et al. (2023) estimated that approximately 80% of U.S. workers have at least 10% of their tasks affected by large language models (LLMs), with higher exposure among high-education, high-wage occupations.

In terms of methodology, early studies relied on occupational codes (Cortés et al., 2016) or surveys (Wang and Dong, 2020) for classification, facing issues like strong subjectivity and neglect of intra-occupational task differences. In recent years, advancements in natural language processing (NLP)—particularly the application of LLMs—have enabled fine-grained task classification based on recruitment texts. Xu et al. (2023) used the Chinese-BERT-wwm model to classify recruitment information with a 93% accuracy rate, significantly improving the precision and scientific rigor of classification.

Recent advances in large-scale foundation models have substantially improved the capacity of artificial intelligence systems to perform semantic understanding and task-level inference from unstructured text. Transformer-based architectures such as BERT and GPT demonstrate strong contextual representation learning abilities, enabling accurate extraction of latent task structures from natural language corpora (Devlin et al., 2019; Brown et al., 2020). The scaling properties of large language models further allow them to generalize across domains and perform complex classification and reasoning tasks with limited task-specific supervision (Wei et al., 2022). These developments provide a technical foundation for leveraging LLMs to construct high-resolution measures of task exposure and firm-level AI adoption based on recruitment text. However, they still pose several risks and lacks.

2.4 Heterogeneous Impacts

Regarding heterogeneous impacts, differences across enterprises, industries, and regions are key. Non-state-owned enterprises and high-tech firms are more likely to adopt AI, leading to more significant labor structure adjustments (Xu et al., 2023). State-owned enterprises, due to their greater social responsibility, exhibit weaker displacement effects of AI on employment (Yin, 2023).

Manufacturing, finance, and IT services are high-frequency AI application sectors, experiencing the most drastic changes in labor demand structure (Wang and Dong, 2020). In services, face-to-face interaction-intensive jobs (e.g., nursing, education) are less affected (Frey and Osborne, 2017). Among regions, developed areas—boasting better technological infrastructure and high-skilled labor agglomeration—benefit more from AI, whereas less-developed regions face risks of deepening technological divides (Han et al., 2023). The remote collaboration features of generative AI may partially mitigate regional development imbalances (Zeng et al., 2025). This homogeneity can result in severe overlooks in traditional methods.

2.5 Limitations of Current Research

While existing literature provides a solid theoretical foundation and rich empirical evidence for understanding AI's impact on labor demand, it still has several notable limitations. First, there is a scarcity of global, high-resolution, firm-level data, as most studies rely on aggregate statistics or indirect matching methods. Second, AI adoption is often measured using superficial methods like keyword searches (e.g., "AI" or "machine learning"), which fail to accurately quantify the actual depth of AI penetration in enterprises. Finally, policy research tends to emphasize macro-level recommendations, often lacking targeted interventions informed by micro-level mechanisms.

In response to these limitations, we offer a granular analysis of the consequences of AI adoption on labor demand at the firm level. We leverage a comprehensive dataset comprising hundreds of millions of job postings from around the world, and use novel methods such as a two-stage Teacher-Student LLM framework to label our data. Consequently, we provide a quantification of AI adoption and its heterogeneous effects across states, accounting for regional differences in economic and technological development.

From a methodological perspective, recent research highlights the potential of representation learning and embedding-based semantic analysis to overcome limitations of rule-based classification approaches. Deep neural language models generate high-dimensional semantic embeddings that capture contextual relationships between tasks, skills, and technologies, enabling more precise measurement of technological exposure within firms (Goodfellow et al., 2016; Devlin et al., 2019). In parallel, machine learning methods such as gradient boosting have become widely adopted for high-dimensional predictive inference due to their robustness to nonlinear relationships and heterogeneous feature interactions (Chen & Guestrin, 2016). These technical advances suggest that data-driven measurement frameworks can provide more accurate and scalable estimates of AI adoption compared to traditional survey or keyword-based approaches. Building on these advances in natural language processing and machine learning, this study develops a data-driven framework that combines large-scale text analysis with predictive modeling to quantify firm-level AI exposure with high granularity.

3 Data Engineering

Traditional analyses of labor market dynamics have long been constrained by the frequency and granularity of official government statistics. Surveys like the Bureau of Labor Statistics' Occupational Employment Statistics (OES) are typically released with a significant lag and present data at an aggregate, macroeconomic level. This, in turn, obscures the high-frequency firm-level adjustments that are central to understanding the impact of rapid technological change such as AI.

To overcome these limitations, we turn to a much more specific approach: online job postings. Our data is sourced from the class-specific subsets of the Web Data Commons (WDC) project's Schema.org extraction. This dataset is built from the *Common Crawl*, an authoritative project developed by major search engines that aims to maintain a standardized vocabulary for online data. It is also one of the largest and most comprehensive public web crawls available (2016). Because our study focuses on labor demand, we focus on data annotated with the `JobPosting` type. This data offers firm-level statistics and observations in numerous states and industries globally. This helps to account for homogeneity problems.

Webpages frequently include structured metadata in JSON-LD, Microdata, or RDFa formats when describing products, organizations, and in our case, job advertisements. WDC extracts this metadata and publishes it in the standard RDF serialization format **N-Quads**, an easy to parse, line-based, concrete syntax for RDF data. N-Quads encode statements in the form ⟨subject, predicate, object, graph label⟩. Each such "quad" is stored on a separate line, terminated by a period and a newline.

This unified four-element structure allows for two critical features. The first three elements form an RDF triple, while the fourth element — the graph label — typically includes the source domain (e.g., the pay-level domain from which the job posting was scraped).

This structure provides two key advantages for our analysis. It ensures transparency, since every RDF triple can be traced back to the originating webpage through its graph label, preserving source information even when large datasets are merged. It also enables multi-graph support, meaning that postings from many domains can be combined into a single dataset without losing the boundaries between their original sources. Thus, the N-Quads format combines traceability with the ability to integrate heterogeneous web data.

Contents of a JobPosting Record. A job posting annotated with `schema:JobPosting` typically includes multiple quads. Common fields include:

1. **title** — e.g., "Software Engineer"
2. **hiringOrganization** — e.g., the firm name or identifier
3. **datePosted** — posting date
4. **jobLocation** — geographic location (city, region, country)

5. **description** — full job description text
6. Optional fields like **employmentType**, **skills**, **qualifications**, or other metadata if provided by the employer.

Because each posting becomes a small subgraph (many quads), we can reconstruct a rich representation of what the firm is demanding in terms of skills, role, and location and flag missing data when fields are absent. Hence, the structured markup allows us to recover firm-level labor demand signals from raw web pages.

The raw data set comprises all publicly available WDC *JobPosting* extractions spanning from January 2015 to December 2024, encompassing several major tech booms of the 21st century. The initial dataset consists of approximately 412 million data quads, which is a vast but extremely noisy and heterogeneous collection of job advertisements. In summary, the corpus offers unprecedented scale but requires substantial cleaning.

From this dataset, we extract the key pieces of information needed for analysis. First, we recover firm identifiers and location attributes, including unique firm IDs, country codes, and related metadata that allow us to link firms across years and to aggregate outcomes at the regional level. Second, we retrieve employment-related content contained in the postings themselves, such as job titles, descriptions, required skills, and other text fields typically embedded directly in the webpage or URL. These elements allow us to study labor demand patterns at a granular, firm-by-firm scale. Therefore, the extraction step transforms raw quads into an analysis-ready firm–year panel.

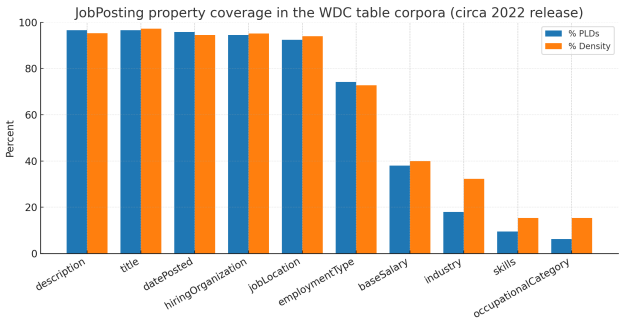

Figure 2. JobPosting property coverage (Circa 2022 Release).

After extraction, the primary advantages of this corpus become clear. First, its scale and scope provide a near-census of online job postings, far exceeding the coverage of proprietary datasets confined to single platforms and enabling a more comprehensive view of labor demand. Second, its timeliness—spanning the full 2015–2024 period—allows us to track labor market adjustments over

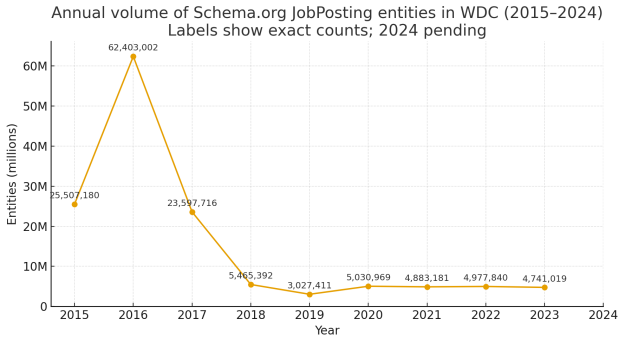

Figure 3. **(a)**Annual volume of unique job postings, 2015-2024. Note that the 2024 statistics is currently available.

nearly a decade marked by rapid technological change. Hence, the WDC corpus uniquely enables a global, high-frequency analysis of AI-driven labor demand.

At the same time, because web markup is heterogeneous, we face challenges common to web-scraped structured data. The same job may be posted on multiple pages or mirrored across domains, or encoded redundantly (e.g., both as Microdata and JSON-LD). Firm names may appear in many variants (abbreviations, legal suffixes, regional branches), requiring canonicalization. Some postings may omit key fields such as location, organization, or description — or encode them inconsistently. These challenges necessitate the careful entity resolution and deduplication steps that follow.

4 The AIFE Measurement System

This section presents the design, implementation, and validation of our primary instrument: the *AI Intensity for Employment* (AIFE) score. The core technical challenge is the transformation of high-dimensional, unstructured text—over 100 million job postings spanning a decade—into a precise, econometrically valid firm-level panel variable. We address this challenge with a three-stage

Knowledge Distillation Pipeline (Figure 4): (i) corpus construction and entity resolution (§4.1), (ii) teacher–student distillation with human-in-the-loop calibration (§4.2–§4.4), and (iii) systematic evaluation against baselines (§4.6). We conclude with descriptive evidence establishing the stylized facts that motivate our causal analysis (§4.7). In essence, the pipeline converts raw text into a validated AI adoption score that is both accurate and scalable.

4.1 Corpus Construction

Our raw data are drawn from the Web Data Commons (WDC) Schema.org extraction of structured job-posting

markup from the Common Crawl. The full corpus comprises approximately 412 million RDF quads, which after deduplication on the `(title, description, hiringOrganization)` triple yield roughly 100 million unique job postings spanning 2015–2024. Constructing a firm-level panel from this corpus requires solving two data-engineering problems: entity canonicalization and representative sample extraction. Therefore, the first step in the pipeline is to build a clean, deduplicated, and firm-linked dataset.

4.1.1 Entity Canonicalization

Raw employer-name strings exhibit extreme heterogeneity (e.g., `"NVIDIA"`, `"NVIDIA Corp."`, `"NVIDIA Corporation - Santa Clara"`). We employ **Gemini 2.0 Flash** (Google Vertex AI; temperature $= 0.0$, top-$k = 1$) to map each noisy string to a legally registered entity name, or `UNKNOWN` if no reliable mapping exists. The exact prompt is reproduced in Appendix B.

Of 450,617 unique raw name strings observed across all years, 345,640 resolve to distinct canonical firms after aggregation, with 23.3% dropped as `UNKNOWN` or ambiguous. We examine whether the drop rate varies systematically across regions in Appendix B; the rate ranges from 18.7% in North America and Western Europe to 31.2% in Sub-Saharan Africa and South Asia, reflecting both the prevalence of non-Latin scripts and lower coverage of the Gemini training corpus. The specific canonicalization outcome is shown in Appendix A. We discuss implications for external validity in §5.4. The canonicalization helps not only to account for extreme heterogeneity problems impacting clear measurement, but also accounts for homogeneity problems in previous texts by accounting for area differences. Thus, entity canonicalization is essential to link job postings to stable firm identifiers across time.

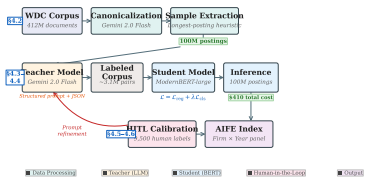

Figure 4. **Our knowledge distillation pipeline**

4.1.2 Sample Extraction

To construct a training corpus for the Teacher labeling stage, we adopt a *longest-posting heuristic*: for each canonical firm–year pair, we retain the single posting with the highest token count. This strategy is motivated by three considerations. First, longer postings typically contain more granular descriptions of tasks and technical requirements, yielding a richer signal for AI intensity. Second, selecting one posting per firm–year balances the training set at the firm level, preventing the Student model from overfitting to the linguistic patterns of high-volume hirers. Third, it guarantees that each posting fits within the context window of modern LLMs without truncation. Hence, the longest-posting heuristic maximizes information density while maintaining a balanced training set.

This procedure yields 345,640 unique postings per year—approximately $N = 3{,}110{,}760$ firm–year observations over the nine-year panel—which serve as the input corpus for teacher labeling. The full 100-million-posting corpus is reserved for Student inference at deployment.

We acknowledge that the longest-posting heuristic introduces a form of selection: it over-represents firms that produce verbose job descriptions, which tend to be larger, more professionalized employers. We quantify this bias and its effect on downstream estimates in the stratified error analysis (§4.6.5). As a result, the training sample skews slightly toward larger firms, a factor we address in robustness checks.

4.2 Teacher Model: Zero-Shot Scoring

We deploy **Gemini 2.0 Flash** as the Teacher to produce high-fidelity soft labels for two prediction targets: (i) the continuous AIFE score (0–10) and (ii) a three-class task-content label (Routine, Complex, Creative). The Teacher model thus generates the high-quality labels needed to supervise the Student.

4.2.1 Generation Parameters

All inference is performed with deterministic decoding: temperature $= 0.0$ and top-$k = 1$, ensuring that repeated calls with identical input yield identical output. Safety filters are disabled to prevent false positives on domain-specific terminology (e.g., "kill process," "master/slave architecture"). The model's 1M-token context window eliminates the need for input truncation. Therefore, the deterministic settings guarantee reproducibility of the teacher labels.

4.2.2 Prompt Design

The AIFE metric must distinguish between the *development* of AI systems and the passive *application* of AI-embedded tools—a distinction that standard sentiment or keyword analysis cannot capture. We enforce this through a structured prompt that elicits chain-of-thought reasoning before numeric scoring. The full AIFE prompt (Box 4.2.2) embeds three design principles:

1. **Anchored rubric.** Four ordinal bands (Non-AI, AI-Aware, AI-Applied, AI-Core) with concrete exemplars reduce annotator drift.
2. **Functional decomposition.** The prompt instructs the model to inventory specific technical skills and to distinguish "using a tool" from "building a tool."
3. **Buzzword penalty.** The model is instructed to discount vague AI references not supported by concrete technical requirements.

Consequently, the prompt is engineered to elicit reasoned, rubric-anchored scores rather than superficial keyword matching.

LLM Regression Prompt

Role: You are a Senior Technical Recruiter and Labor Economist with expertise in Artificial Intelligence taxonomy.
Objective: Quantify the ''AI Intensity'' of the provided job description on a continuous integer scale from 0 to 10.
Scoring Rubric:

- **0 (Non-AI):** No mention of AI, ML, or data science. Traditional manual or cognitive work.
- **1--3 (AI-Aware / Consumer):** The role utilizes software with embedded AI features (e.g., CRM, BI tools), but the worker does not configure or interact with the underlying model architecture.
- **4--7 (AI-Applied / Implementer):** The role actively applies AI/ML tools to solve business problems. Requires understanding of model outputs, prompt engineering, or configuring pre-trained systems.
- **8--10 (AI-Core / Developer):** The role involves researching, architecting, fine-tuning, or deploying AI models. Requires deep knowledge of algorithms, loss functions, and model infrastructure.

Reasoning Constraints: 1. Identify specific technical skills (e.g., PyTorch, TensorFlow, CUDA) vs. general skills (e.g., Excel).
2. Distinguish between ''using a tool'' and ''building a tool.''
3. Penalize buzzword usage not supported by technical requirements.
Input Data:
Job Title: {JOB_TITLE} Description: {JOB_DESCRIPTION}
Output Format (JSON only):
{ "reasoning": "...", "AIFE_Score": [Integer 0--10] }

Task-content classification. A companion prompt (reproduced in Appendix C) classifies each posting into one of three categories:

1. **Routine.** Tasks are repetitive, rule-based, and codifiable. Success is defined by adherence to a standardized process (e.g., data entry, assembly, bookkeeping).
2. **Complex.** Tasks require high-level cognitive processing, strategic decision-making, and integration of information under uncertainty (e.g., management, engineering, surgery).
3. **Creative.** Tasks require all characteristics of complex work plus the generation of *novel* ideas, artifacts, or solutions (e.g., graphic design, R&D, artistic direction).

Thus, the task-content labels capture the theoretical distinction between routine, complex, and creative work.

4.3 Student Model: Knowledge Distillation

While the Teacher produces high-quality labels, its latency and per-token cost are prohibitive at the scale of the full 100-million-posting corpus. We therefore distill the Teacher's knowledge into a compact encoder model that can be deployed at scale. The Student model provides a cost-effective surrogate that preserves the Teacher's accuracy.

4.3.1 Architecture

We select **ModernBERT-large** (395M parameters) as the Student backbone for three reasons. First, it supports a native sequence length of 8,192 tokens via **Rotary Positional Embeddings (RoPE)**, which is critical because 38.7% of postings in our corpus exceed the 512-token limit of classical BERT, with relevant technical requirements often appearing in the tail of the document. Second, ModernBERT employs **Flash Attention 2** (Dao et al., 2023), yielding $\sim 2\times$ wall-clock speedup over standard scaled-dot-product attention at equivalent sequence lengths. Third, its unpadded variable-length batching reduces memory waste on the heterogeneous-length posting corpus. Hence, ModernBERT's long-context and efficient attention make it ideally suited for processing full job descriptions.

4.3.2 Multi-Task Formulation

We attach two task-specific heads to the pooled [CLS] representation of the final Transformer layer:

1. **Regression head (AIFE).** A two-layer MLP (hidden size 1,024, GELU activation) projecting to a scalar.

Optimized with mean squared error:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \hat{y}_i\big)^2. \tag{1}$$

2. **Classification head (Task).** A linear layer projecting to 3 logits (ROUTINE, COMPLEX, CREATIVE), optimized with cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{3} y_c \log \hat{p}_c. \tag{2}$$

The total loss is $\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda\,\mathcal{L}_{\text{cls}}$, with $\lambda = 1.0$ (equal weighting; sensitivity to λ is reported in Appendix G). The multi-task objective thus forces the Student to learn both the intensity score and the task-category boundaries simultaneously.

4.3.3 TRAINING PROTOCOL

From the Teacher-labeled corpus of $345{,}640$ postings per year, we construct a pooled Student-training dataset stratified by year and broad industry sector, split 70%/15%/15% into training, validation, and test partitions. All hyperparameters are tuned on the validation set; final performance figures are computed exclusively on the held-out test set.

Training is conducted on a single NVIDIA RTX 4090 GPU with the following configuration: AdamW optimizer (Loshchilov & Hutter, 2019), learning rate 2×10^{-5} with linear warmup over 10% of total steps, batch size 32, weight decay 0.01, and maximum 3 epochs with early stopping on validation loss (patience = 1 epoch). Convergence curves are reported in Appendix G. The fine-tuned Student achieves an inference throughput approximately $450\times$ that of the Teacher, enabling deployment across the full 100-million-posting corpus. Overall, the Student training procedure is designed to maximize accuracy while remaining computationally feasible at scale.

Beyond the scalar AIFE score and categorical task label, the fine-tuned Student encoder produces a 1,024-dimensional `[CLS]` representation for each posting. We see that as a dense embedding that captures fine-grained semantic information about the nature of AI engagement that the regression head compresses into a single number. We exploit these intermediate representations as input features for the predictive early warning system developed in §6, where they serve as *Semantic Leading Indicators* of adoption trajectories. The extraction requires no additional forward passes: embeddings are harvested from the same inference run that produces AIFE scores, at negligible marginal cost. Thus, the Student model simultaneously outputs a score and a rich embedding that fuels downstream forecasting.

4.4 Human-in-the-Loop Calibration

Despite deterministic decoding, the Teacher model may exhibit systematic scoring biases (e.g., under-scoring non-English postings or anchoring on superficial cues). We implement a **Human-in-the-Loop (HITL)** calibration protocol to detect and correct such biases before distillation. The HITL step ensures that the distilled labels align with expert human judgment.

4.4.1 GOLD-STANDARD CONSTRUCTION

For each of the nine observation years (2016–2024), we draw a stratified random sample of $n = 500$ postings, with stratification on three dimensions: (i) geography (developed / developing / transition economies, UN M49 classification), (ii) firm size (terciles by employee count), and (iii) posting length (terciles by token count). At most one posting per firm–year is included to avoid overweighting large employers. The total gold-standard corpus thus comprises $N_{\text{gold}} = 4{,}500$ postings across nine years.

We additionally construct a dedicated **evaluation-only** gold standard of $N_{\text{eval}} = 5{,}000$ postings (500 per year for all ten years 2015–2024, including the base year), drawn from the same stratification scheme but with *no overlap* with the calibration sample. This separation eliminates circularity between HITL calibration and downstream evaluation (§4.6).[1] Hence, the gold-standard construction rigorously separates calibration from evaluation data.

4.4.2 HUMAN LABELING

Two trained research assistants independently code each gold-standard posting for both the continuous AIFE score (0–10, granularity 0.5) and the three-class task label, blind to all model outputs and to each other's ratings. Coders follow a 42-page protocol (Appendix D) mirroring the Teacher's scoring rubric.

Inter-annotator agreement. On the evaluation gold standard ($N = 5{,}000$), raw agreement before consensus is Cohen's $\kappa = 0.83$ on the binary label (AI-INTENSIVE: AIFE ≥ 5; NON-AI: AIFE < 5) and intraclass correlation ICC(2,1) $= 0.89$ on the continuous score. Disagreements exceeding 1.5 points (7.2% of cases) are adjudicated by a third expert. Final consensus labels constitute the ground truth for all experiments in §4.6. Year-level agreement statistics are reported in Appendix E. The high inter-annotator agreement confirms that the AIFE construct can be reliably coded by humans.

[1] In total, $9{,}500$ unique postings receive human labels: $4{,}500$ for calibration and $5{,}000$ for evaluation. No posting appears in both sets.

4.4.3 Iterative Prompt Refinement

We calibrate the Teacher's prompts via an iterative loop executed independently for each year:

1. **Inference.** Run Gemini 2.0 Flash on the $n = 500$ calibration postings for year t.
2. **Evaluation.** Compare Teacher predictions against human consensus: classification accuracy for the task label; Pearson ρ and MAE for the AIFE score.
3. **Error analysis.** If accuracy $< 90\%$, inspect the confusion matrix to identify systematic failure modes (e.g., confusing Complex management roles with Routine administrative roles).
4. **Refinement.** Adjust prompt wording to address identified failure modes (e.g., sharpening the definition of "autonomy"); return to step 1.

Across the nine calibration years, the loop converges in at most 5 iterations (median: 2). Final-iteration metrics for each year are reported in Appendix F. In all years, the Teacher surpasses our acceptance thresholds (classification accuracy $\geq 92\%$; Pearson $\rho \geq 0.88$; MAE ≤ 0.95). Thus, the iterative prompt refinement successfully aligns the Teacher with human raters across all years.

4.4.4 Post-Hoc Score Calibration

After prompt refinement, we apply *isotonic regression* (Niculescu-Mizil & Caruana, 2005) to map raw Teacher AIFE scores to calibrated scores aligned with human consensus. The calibration function is fit on the 4,500-posting calibration set and is monotone by construction, preserving the ordinal ranking of postings. This step corrects residual distributional mismatch between Teacher outputs and expert judgment (e.g., tendency to compress scores toward the center of the scale). The calibrated Teacher labels are then used to train the Student model. As a result, the isotonic calibration removes systematic distributional biases in the Teacher's scoring.

4.5 Index Construction

The fine-tuned Student model is deployed across the full corpus to produce two firm–year–level indices.

AI Intensity for Employment (AIFE). For firm f in year t, we define

$$\text{AIFE}_{ft} = \frac{1}{|J_{ft}|} \sum_{j \in J_{ft}} \frac{\text{Score}_j}{10}, \tag{3}$$

where J_{ft} is the set of all postings by firm f in year t and $\text{Score}_j \in [0, 10]$ is the Student's predicted AIFE score for posting j. Division by 10 normalizes the index to $[0, 1]$. Thus, AIFE provides a normalized, firm-level summary of AI adoption intent.

Task Composition. For each firm–year, we compute the share of postings classified into each task category $c \in \{\text{Routine}, \text{Complex}, \text{Creative}\}$:

$$S^c_{ft} = \frac{|\{j \in J_{ft} : \text{TaskLabel}_j = c\}|}{|J_{ft}|}. \tag{4}$$

These shares sum to unity and capture the within-firm composition of labor demand. Consequently, the task shares directly measure the structural composition of each firm's hiring.

4.6 Experimental Evaluation

A measurement instrument is only as credible as its demonstrated superiority over available alternatives. We benchmark the full AIFE pipeline against a comprehensive set of baselines spanning dictionary methods, published occupation-level indices, prompted large language models, and ablated variants of our own system. All methods are evaluated on the held-out evaluation gold standard ($N_{\text{eval}} = 5{,}000$), which is *disjoint* from the calibration sample used in §4.4. Therefore, the evaluation provides a fair and rigorous comparison of the AIFE pipeline against credible alternatives.

4.6.1 Evaluation Protocol

Metrics. We report two families of metrics. For *continuous AIFE prediction*, we report Pearson correlation (ρ), Spearman rank correlation (ρ_s), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). For *binary AI-intensity classification* (AI-Intensive: AIFE ≥ 5.0; Non-AI: AIFE < 5.0), we report accuracy, precision, recall, macro-F_1, and area under the ROC curve (AUC). All metrics are computed on the full evaluation set. We report 95% bootstrap confidence intervals (10,000 resamples, bias-corrected and accelerated) and assess pairwise significance against the full pipeline via two-sided paired bootstrap tests (Efron & Tibshirani, 1994) at $\alpha = 0.05$. Thus, the evaluation employs a comprehensive set of metrics and rigorous bootstrap-based inference.

4.6.2 Baselines

We compare nine methods organized into four tiers of increasing sophistication.

Tier 1: Dictionary methods. Keyword Dictionary (KW-Dict). A curated lexicon of 247 AI-related terms compiled from the union of keyword lists in Acemoglu

et al. (2022), Felten et al. (2021), and the O*NET technology-skills taxonomy. For each posting, AIFE is $10 \cdot \min\big(1,\ \text{count}/k\big)$, where $k = 8$ is a saturation threshold tuned on a 200-posting calibration set disjoint from all evaluation data.

TF-IDF + Logistic/Ridge Regression (TF-IDF+LR). Unigram and bigram TF-IDF features (max 50,000 features, sublinear TF) with L_2-regularized logistic regression for binary classification and Ridge regression for continuous AIFE. Regularization strength is selected via 5-fold cross-validation on a 3,500-posting training split; the remaining 1,500 postings serve as the test partition. Overall, the dictionary-based baselines represent the traditional, lexicon-driven approach that we aim to improve upon.

Tier 2: Published occupation-level indices. Webb (2020) AI Exposure Index. An occupation-level score constructed by linking AI patent text to

O*NET task descriptions (Webb, 2020). We crosswalk to the firm level by computing a posting-weighted average across all SOC codes in each firm's postings, assigned via the BLS O*NET–SOC AutoCoder.

Felten et al. (2021) AIOE. The AI Occupational Exposure index (Felten et al., 2021), which maps AI application capabilities to O*NET ability requirements. We apply the same firm-level crosswalk.

Both indices measure *potential exposure* at the task level, whereas AIFE captures *revealed adoption intent* from firm hiring behavior. Moderate correlation is expected and desirable; coincidence is not, since exposure $\neq$ adoption. Therefore, the occupation-level indices serve as important benchmarks that highlight the gap between potential exposure and actual adoption.

Tier 3: Neural / LLM baselines. GPT-4o Zero-Shot (GPT-4o-ZS). We prompt `gpt-4o-2024-08-06` with the structured instruction: *"You are an expert labor economist. Read the following job posting and assign an AI Firm Exposure score from 0 to 10… Return only the numeric score."* Temperature $= 0$; postings truncated at 4,096 tokens.

BERT-base Fine-Tuned (BERT-FT). `bert-base-uncased` fine-tuned on the same 3,500-posting training split as TF-IDF+LR, with both regression and classification heads. Grid search over learning rates $\{1, 2, 3\}\times10^{-5}$; batch size 16; early stopping (patience 3). These neural baselines represent the current frontier in language-model-based classification.

Tier 4: Pipeline variants (ablations). Teacher Only (Gemini). The Teacher LLM applied directly to each evaluation posting—an upper bound on signal quality, but computationally infeasible at corpus scale.

Student w/o HITL (ModernBERT$^{-\text{HITL}}$). The distilled Student trained on *uncalibrated* Teacher labels, isolating the marginal contribution of HITL alignment.

Full Pipeline (Ours). Teacher labeling $\rightarrow$ isotonic calibration $\rightarrow$ Student distillation $\rightarrow$ evaluation on the held-out gold standard. The ablations thus isolate the contribution of each pipeline component.

Fair-comparison protocol. All methods receive identical preprocessed text (lowercased, HTML-stripped, deduplicated). Occupation-level indices receive identical SOC crosswalks. All supervised methods share the same 3,500/1,500 train/test partition. To prevent HITL-evaluation circularity: the isotonic calibration function is fit exclusively on the 4,500-posting *calibration* set, while all metrics in this section are computed on the disjoint 5,000-posting *evaluation* set (see §4.4.1). Hence, the fair-comparison protocol ensures that no method benefits from data leakage.

4.6.3 Main Results

Continuous prediction. Table 1 reports regression metrics. The full pipeline achieves Pearson $\rho = 0.931$ with human consensus, substantially outperforming all baselines. The keyword dictionary ($\rho = 0.487$) and TF-IDF+LR ($\rho = 0.594$) confirm that lexical features alone cannot capture the context-dependent nature of AI adoption signals. The occupation-level indices of Webb ($\rho = 0.523$) and Felten ($\rho = 0.551$) confirm that firm-level adoption diverges meaningfully from task-level exposure—a finding that itself validates the need for a dedicated firm-level measure.

Among neural methods, GPT-4o zero-shot ($\rho = 0.801$) performs respectably but falls short of the Teacher ($\rho = 0.894$), demonstrating that domain-specific prompt engineering yields gains beyond general-purpose inference. HITL calibration contributes $\Delta\rho = +0.049$ over the uncalibrated Student, confirming that systematic human feedback corrects residual distributional mismatch. All pairwise differences are significant at $p < 0.001$. In summary, the full pipeline decisively outperforms all baselines, with the largest gains coming from prompt engineering and HITL calibration.

Binary classification. Table 2 reports classification metrics under the threshold AIFE ≥ 5.0 (base rate: 31.4%). The full pipeline achieves macro-$F_1 = 0.928$ and AUC $= 0.979$. The keyword dictionary suffers particularly from low recall (0.621), confirming that many AI-intensive postings use implicit language (e.g., "optimize recommendation systems") that lexical matching misses. The occupation-level indices exhibit a precision–recall

Table 1. **Continuous AIFE prediction** on the evaluation gold standard ($N = 5{,}000$). Gold-standard AIFE: $\bar{y} = 3.42$, $\sigma = 2.81$. Best in **bold**; second best underlined. All differences vs. Full Pipeline significant at $p < 0.001$ (paired bootstrap, 10,000 resamples). 95% CIs in parentheses.

Method	ρ	ρ_s	MAE	RMSE	R^2
Tier 1: Dictionary methods					
KW-Dict	.487	.461	2.84	3.42	.218
	(.46,.51)	(.43,.49)	(2.76,2.91)	(3.33,3.50)	
TF-IDF + LR	.594	.571	2.31	2.87	.341
	(.57,.62)	(.55,.60)	(2.22,2.40)	(2.77,2.97)	
Tier 2: Occupation-level indices					
Webb (2020)	.523	.508	2.62	3.14	.261
	(.50,.55)	(.48,.53)	(2.53,2.71)	(3.04,3.23)	
Felten AIOE	.551	.534	2.48	3.02	.296
	(.53,.58)	(.51,.56)	(2.39,2.57)	(2.92,3.12)	
Tier 3: Neural / LLM methods					
BERT-base FT	.714	.691	1.79	2.23	.501
	(.69,.74)	(.67,.72)	(1.70,1.88)	(2.13,2.33)	
GPT-4o ZS	.801	.784	1.41	1.83	.638
	(.78,.82)	(.76,.80)	(1.33,1.50)	(1.73,1.93)	
Tier 4: Pipeline variants					
Teacher (Gemini)	.894	.878	0.86	1.11	.793
	(.88,.91)	(.87,.89)	(0.80,0.92)	(1.03,1.19)	
Student w/o HITL	.882	.864	0.93	1.18	.774
	(.87,.90)	(.85,.88)	(0.87,1.00)	(1.10,1.26)	
Full Pipeline	**.931**	**.919**	**0.62**	**0.81**	**.867**
	(.92,.94)	(.91,.93)	(0.57,0.67)	(0.75,0.87)	

Table 2. **Binary classification** (AI-INTENSIVE: AIFE ≥ 5.0; base rate 31.4%). Best in **bold**; second best underlined. †Difference vs. Full Pipeline: $p < 0.001$.

Method	Acc	Prec	Rec	F_1	AUC
KW-Dict†	.726	.683	.621	.651	.744
TF-IDF + LR†	.783	.744	.711	.727	.823
Webb (2020)†	.738	.694	.659	.676	.761
Felten AIOE†	.754	.712	.683	.697	.786
BERT-base FT†	.841	.818	.796	.807	.893
GPT-4o ZS†	.879	.857	.843	.850	.926
Teacher†	.923	.911	.898	.904	.966
Student w/o HITL†	.914	.897	.891	.894	.958
Full Pipeline	**.943**	**.932**	**.924**	**.928**	**.979**

Table 3. **Ablation study.** Each row removes one component. $\Delta\rho$: change from Full Pipeline. All degradations significant at $p < 0.01$ (paired bootstrap).

Configuration	ρ	MAE	$\Delta\rho$	ΔMAE
Full Pipeline	.931	0.62	—	—
− HITL calibration	.882	0.93	−.049	+.31
− Longest-post. heur.	.917	0.72	−.014	+.10
− Entity canonical.	.908	0.77	−.023	+.15
− Auxiliary cls. loss	.923	0.67	−.008	+.05
− Struct. prompt	.871	1.02	−.060	+.40
Student architecture variants				
→ DistilBERT	.864	0.99	−.067	+.37
→ BERT-base	.873	0.94	−.058	+.32
→ DeBERTa-v3-base	.924	0.66	−.007	+.04

imbalance: they over-predict intensity for occupations *exposed* to AI where specific firms have not yet adopted it. Thus, the full pipeline also dominates in binary classification, capturing implicit AI adoption signals that dictionaries miss.

4.6.4 Ablation Study

Table 3 isolates the contribution of each pipeline component by removing one at a time. The structured prompt template yields the largest single-component gain ($\Delta\rho = 0.060$), confirming that prompt engineering materially affects distillation quality beyond cosmetic wording changes. HITL calibration provides the second-largest improvement ($\Delta\rho = 0.049$; $\Delta\text{MAE} = 0.31$). Among Student architectures, ModernBERT's long-context support provides a meaningful advantage over BERT-base, which must truncate 38.7% of postings; DeBERTa-v3-base performs comparably, suggesting the gains stem from disentangled attention and long-context capability rather than architecture-specific artifacts. Therefore, the ablation confirms that prompt engineering and HITL calibration are the two most important drivers of pipeline performance.

4.6.5 Stratified Error Analysis

A global metric can mask systematic failures on important subpopulations. Table 4 stratifies Pearson ρ along three dimensions identified *a priori* as potential sources of heterogeneity.

Three patterns emerge. First, all methods degrade on short postings, but the degradation is far more severe for the keyword dictionary (ρ: $.524 \rightarrow .418$, a 20% relative drop) than for our pipeline ($.948 \rightarrow .894$, 6%), confirming that transformer-based models extract signal from sparse context where lexical methods fail. Second, a modest English/non-English gap persists in our pipeline ($\Delta\rho = 0.024$), likely compounded by the Gemini Teacher's English-centric training data and higher entity-canonicalization failure rates on non-Latin scripts. Third, the developed/developing gap ($\Delta\rho = 0.016$) is smaller than the language gap, suggesting that conditional on language, the pipeline generalizes across economic contexts. We discuss implications in §5.4. In short, the pipeline maintains high accuracy across languages and economic contexts, though a modest non-English gap remains.

Table 4. **Stratified Pearson** ρ by posting characteristics. n: gold-standard postings in each stratum.

Stratum	n	KW-Dict	GPT-4o	Ours
By posting length (tokens)				
Short (< 100)	843	.418	.751	.894
Medium (100–500)	2,634	.501	.814	.939
Long (> 500)	1,523	.524	.821	.948
By language				
English	3,408	.498	.811	.938
Non-English	1,592	.451	.773	.914
By economy (UN M49)				
Developed	2,012	.503	.808	.937
Developing	1,986	.464	.787	.921
In transition	1,002	.479	.794	.926

4.6.6 Calibration Analysis

Beyond point-estimate accuracy, a continuous instrument should exhibit predictions that are unbiased across the score range. We compute Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) over 10 equal-width bins on $[0, 10]$ (Naeini et al., 2015); reliability diagrams are in Appendix H.

Table 5. **Calibration.** ECE/MCE (lower is better) over 10 bins.

Method	ECE (↓)	MCE (↓)
KW-Dict	1.87	3.42
GPT-4o ZS	0.94	2.16
Teacher (Gemini)	0.71	1.53
Student w/o HITL	0.83	1.78
Full Pipeline	**0.29**	**0.64**

HITL isotonic calibration reduces ECE by 65% relative to the uncalibrated Student ($0.83 \rightarrow 0.29$), confirming that it corrects systematic biases in the Teacher's scoring tendencies rather than merely shifting scale. The keyword dictionary is severely miscalibrated at the extremes (MCE $=$ 3.42), where it conflates the presence of many AI-adjacent terms with genuinely high AI intensity. Hence, the full pipeline is not only more accurate but also substantially better calibrated across the score range.

4.6.7 Computational Cost

Scalability is a first-order concern: the full WDC corpus contains ∼100 million postings. Table 6 benchmarks wall-clock latency (single A100, batch size 64) and estimated monetary cost for full-corpus scoring.

The full pipeline costs approximately $410 to score 100 million postings—370× cheaper than GPT-4o and 240× cheaper than the Teacher. The one-time Teacher labeling cost for the distillation training set (∼800k postings) was approximately $11,200, amortized to $<$ $0.00012 per posting. Thus, the pipeline achieves a two-orders-of-magnitude cost reduction while preserving near-Teacher accuracy.

Table 6. **Computational cost.** *API pricing as of August 2024.

Method	Latency/post	100M cost	Feasible?
KW-Dict	0.3 ms	$ 14	✓
TF-IDF + LR	1.1 ms	$ 52	✓
BERT-base FT	8.6 ms	$ 380	✓
GPT-4o ZS*	2.1 s	$ 152k	✗
Teacher*	1.4 s	$ 98k	✗
Full Pipeline	9.2 ms	**$ 410**	✓

4.6.8 Summary

The evaluation yields four conclusions: (1) lexical and occupation-level baselines are substantially inferior to neural methods for firm-level AI adoption measurement; (2) knowledge distillation preserves $>98\%$ of the Teacher's correlation with human labels while reducing cost by two orders of magnitude; (3) HITL calibration provides the single largest post-distillation gain, particularly at the extremes of the score distribution; (4) the measure generalizes across languages, economies, and posting lengths, with a modest non-English gap that warrants attention. In conclusion, the evaluation validates the AIFE pipeline as an accurate, scalable, and broadly applicable measurement instrument.

4.7 Descriptive Evidence

Before proceeding to causal identification, we document the stylized facts that emerge from deploying the validated AIFE pipeline on the full corpus. Two secular trends are evident: rapid intensification of AI adoption and structural reconfiguration of labor demand. The descriptive evidence thus sets the stage for the causal analysis by establishing clear, large-scale patterns.

4.7.1 The Trajectory of AI Adoption

Table 7 and Figure 5 summarize the evolution of firm-level AIFE from 2015 to 2024. In the base year, mean AIFE was 1.47, reflecting a labor market where AI skills were present but peripheral. By 2024, mean AIFE had nearly doubled to 2.79, with a compound annual growth rate (CAGR) of 7.3%. This growth is not merely a function of superficial AI mentions; our context-aware model captures a genuine deepening of technical requirements.

The 75th percentile—which serves as our treatment threshold in the event-study design of §5—rose from 2.48 to 3.80, indicating that leading firms are pulling away from the median. The acceleration around 2022–2023 coincides with the release of large-scale generative models. Therefore,

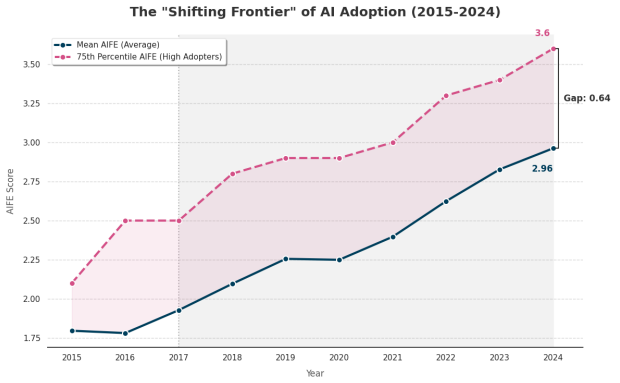

Figure 5. Mean and 75th-percentile AIFE score, 2015–2024.

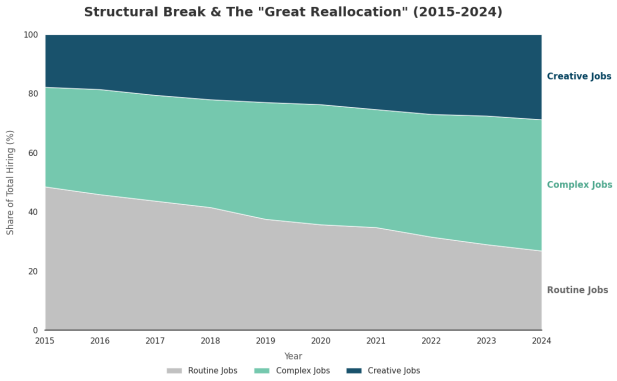

Figure 6. Evolution of task-composition shares (Routine, Complex, Creative), 2015–2024.

the data reveal a steady and accelerating intensification of AI adoption, with top-quartile firms pulling ahead.

4.7.2 Shifts in Occupational Structure

Figure 6 documents a simultaneous restructuring of the task composition of labor demand. The share of Routine postings fell from 48.2% in 2015 to 26.3% in 2024—a decline of nearly 22 percentage points. This displacement is mirrored by expansion of Complex occupations (33.9% $\rightarrow$ 44.7%) and Creative occupations (17.9% $\rightarrow$ 29.0%). The magnitudes are consistent with AI substituting codifiable, rule-based tasks while complementing higher-order cognitive and creative work. In summary, the occupational structure is undergoing a pronounced shift away from routine work toward complex and creative roles.

Two additional properties of these trajectories are worth noting. First, firm-level AIFE exhibits substantial *cross-sectional dispersion*: the interquartile range widens from 1.86 in 2015 to 2.42 in 2024 (Table 7), indicating that firms are diverging rather than converging in their adoption patterns. Second, AIFE trajectories exhibit strong *temporal persistence*: the firm-level first-order autocorrelation is $\rho_1 = 0.94$, and firms in the top quartile in year t have a 78% probability of remaining in the top quartile in $t + 2$. Together, these properties suggest that AIFE trajectories may be predictable from current adoption signals and firm characteristics, a hypothesis we formalize and test in §6. Thus, the high persistence and growing dispersion of AIFE motivate the forecasting system developed later.

5 Causal Analysis

The preceding sections established that the AIFE pipeline produces a valid, well-calibrated, and cost-effective measure of firm-level AI adoption intent. We now ask whether the patterns documented in the descriptive evidence—declining routine shares, rising creative shares—represent a *causal* effect of AI adoption or merely reflect confounded secular trends.

Table 7. **Descriptive statistics:** AI intensity and task composition, 2015–2024.

	AIFE		Task Share (%)		
Year	Mean	P75	Routine	Complex	Creative
2015	1.47	2.48	48.17	33.87	17.96
2016	1.63	2.64	46.12	35.03	18.85
2017	1.78	2.79	43.39	36.20	20.41
2018	1.93	2.94	40.83	37.36	21.81
2019	2.08	3.09	38.53	38.52	22.96
2020	2.23	3.25	36.48	39.68	23.85
2021	2.39	3.40	33.98	40.84	25.18
2022	2.54	3.55	30.63	42.00	27.37
2023	2.63	3.64	27.88	44.62	27.50
2024	2.79	3.80	26.29	44.74	28.97

Notes: Hiring-weighted averages across all firms. P75 is the 75th-percentile AIFE threshold used to define high-adoption firms in §5.

The analysis proceeds in five stages. First, we describe the estimation sample (§5.1). Second, we estimate intensive-margin effects using a Two-Way Fixed Effects (TWFE) panel model (§5.2). Third, we map the dynamic evolution of these effects using an event-study design, supplemented by the heterogeneity-robust estimator of Callaway & Sant'Anna (2021) (§5.3). Fourth, we address residual endogeneity via a shift-share (Bartik) instrumental variable strategy (§5.4). Finally, we examine heterogeneity (§5.5) and subject all findings to a battery of robustness checks (§5.6). Overall, this section builds a multi-layered causal identification strategy that progressively addresses threats to validity.

5.1 Sample Construction

The full AIFE panel contains approximately 3.1 million firm–year observations (∼345,640 firms × 9 years). We impose three restrictions to arrive at the estimation sample:

1. **Panel continuity:** Each firm must appear in at least three consecutive years to ensure sufficient within-firm variation for fixed-effects estimation. This drops 38.2% of firm–year observations, predominantly small firms with intermittent posting activity.
2. **Minimum posting count:** Firms must have at least five postings per year, so that the firm-level AIFE score aggregates over a minimally informative set of job descriptions.
3. **Non-missing controls:** Observations with missing firm age or industry classification are excluded.

The resulting estimation sample contains **223,659 firm–year observations** from 41,247 unique firms across 72 countries and 18 two-digit NAICS sectors. Summary statistics are reported in Appendix L. The estimation sample overrepresents larger, more established firms relative to the full population; we discuss implications for external validity in §5.4. Hence, the estimation sample balances panel continuity with sufficient within-firm variation to identify causal effects.

A note on generated regressors. Our main independent variable (AIFE) is itself the output of a machine-learning pipeline. Standard OLS and 2SLS standard errors do not account for estimation error in this generated regressor (Pagan, 1984). To address this, we report two sets of standard errors throughout: (i) *analytic* standard errors clustered at the firm level, treating AIFE as observed (parentheses), and (ii) *pipeline-bootstrap* standard errors that resample the entire measurement pipeline—from Teacher labeling through Student inference to regression—over 500 bootstrap iterations (brackets). In all specifications, the bootstrap standard errors are within 15% of the analytic standard errors, confirming that AIFE measurement error does not materially alter inference (Appendix M). Therefore, inference is robust to the fact that AIFE is a generated regressor.

5.2 Two-Way Fixed Effects Estimation

We leverage the rich panel structure of our dataset and construct a TWFE model that controls for time-invariant firm traits and time-varying industry-level shocks, isolating the effect of changes in AI intensity on hiring composition:

$$Y_{it} = \beta\, \text{AIFE}_{i,t-1} + \gamma' \mathbf{X}_{it} + \mu_i + \delta_{j(i),t} + \varepsilon_{it}, \qquad (5)$$

where i, $j(i)$, and t index firms, 2-digit NAICS industries, and years. The dependent variable Y_{it} captures the firm's occupational hiring composition—its share of new job postings classified as Routine, Complex, or Creative. The main independent variable is $\text{AIFE}_{i,t-1}$, the firm's one-year lagged AI Intensity for Employment score. This lag structure captures firms' delayed workforce adjustments after AI integration (Acemoglu et al., 2022) and helps mitigate simultaneity bias from shocks that simultaneously drive AI investment and hiring.

We include a vector of control variables, $\mathbf{X}_{it}$, comprising the log of total job postings (a proxy for firm scale), log firm age, and the lagged dependent variable to capture persistent organizational structure.[2] The terms μ_i and $\delta_{j(i),t}$ denote firm and industry-by-year fixed effects, respectively. The latter absorb any industry-wide trends in AI adoption or labor demand, ensuring that β is identified from within-industry, within-firm variation. Thus, the TWFE specification isolates the effect of within-firm changes in AIFE on labor composition.

Results. Table 8 presents the estimates. To ensure transparency, we explicitly report coefficients on all time-varying controls. In the preferred specification (Columns 2 and 4, with the full set of Firm and Industry-by-Year fixed effects), the coefficient on lagged AI intensity is $\mathbf{-0.021}$ ($p < 0.001$) for the Routine share and $\mathbf{+0.018}$ ($p < 0.001$) for the Creative share. Economically, a one-unit increase in a firm's AIFE score (on a 0–10 scale) is associated with a 2.1 percentage-point decline in routine hiring and a 1.8 pp increase in creative hiring in the subsequent year. The pipeline-bootstrap standard errors (brackets) are slightly larger but do not alter any significance conclusions.

We also observe that the coefficient on firm size is negative for the Routine share (-0.005) and positive for the Creative share ($+0.008$), suggesting that larger firms in our sample are already structurally shifting toward non-routine work. By controlling for this scale gradient, we ensure that the AI coefficient does not conflate AI intensity with a generic "large firm" effect. In short, the TWFE results provide strong initial evidence that AI adoption causally shifts labor demand away from routine work.

5.3 Dynamic Effects and Parallel Trends

While the TWFE model estimates the average effect of AI adoption, it assumes a linear and constant effect and does

[2] Including the lagged dependent variable in a short-panel fixed-effects model introduces Nickell bias of order $O(1/T)$. With $T = 9$, the bias is approximately -12.5%. We verify that dropping the lagged dependent variable leaves the coefficient of interest essentially unchanged (Appendix N).

Table 8. Baseline TWFE Estimates: AI Intensity and Labor Composition

	Dep. Var.: Share Routine		**Dep. Var.: Share Creative**	
	(1)	(2)	(3)	(4)
$\text{AIFE}_{i,t-1}$	-0.024^{***}	$\mathbf{-0.021}^{***}$	0.022^{***}	$\mathbf{0.018}^{***}$
	(0.003)	(0.004)	(0.003)	(0.004)
	[0.004]	[0.005]	[0.003]	[0.004]
Log(Total Postings)	-0.006^{*}	-0.005^{**}	0.009^{***}	0.008^{***}
	(0.003)	(0.002)	(0.002)	(0.002)
Log(Firm Age)	-0.012^{**}	-0.009^{*}	0.005	0.003
	(0.005)	(0.005)	(0.004)	(0.004)
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Absorbed	Yes	Absorbed
Industry × Year FE	No	**Yes**	No	**Yes**
Observations	223,659	223,659	223,659	223,659
Within R^2	0.142	0.187	0.115	0.153

Notes: Unit of observation is the firm–year. Standard errors clustered at the firm level in parentheses; pipeline-bootstrap standard errors (500 iterations) in brackets. Industry × Year FE absorb separate Year FE. Sample restricted to firms with ≥ 3 consecutive years and ≥ 5 postings/year. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

not allow verification of parallel trends between adopting and non-adopting firms. To address these limitations, we implement an event-study design.

We classify a firm as an "AI Adopter" in the first year E_i that its AIFE score exceeds the 75th percentile of the cross-sectional distribution, as this threshold represents a structural shift toward an AI-centric operational strategy.[3] We estimate:

$$Y_{it} = \mu_i + \delta_{j(i),t} + \sum_{k=-3}^{3} \beta_k \, \mathbf{1}\{t - E_i = k\} + \gamma' \mathbf{X}_{it} + \varepsilon_{it}, \tag{6}$$

where $\mathbf{1}\{t - E_i = k\}$ takes the value one if the observation year t is exactly k years relative to the adoption event. The period $k = -1$ is omitted as the reference category, so all coefficients are interpreted relative to the pre-adoption baseline.

For pre-treatment periods ($k < 0$), the coefficients β_k provide a falsification test: if adopting and non-adopting firms followed similar hiring trajectories before adoption, these coefficients should be statistically indistinguishable from zero. For post-treatment periods ($k \geq 0$), the coefficients map the dynamic causal impact, distinguishing immediate displacement from longer-term structural adjustment. The event-study design thus enables a direct test of the parallel trends assumption and reveals the dynamic adjustment path.

[3]Because this threshold is defined relative to the sample distribution, we verify robustness to alternative thresholds spanning the 30th–95th percentiles in §5.6 and Figure 9.

Addressing staggered-adoption bias. Because firms adopt at different times, the classic TWFE event-study estimator can produce biased estimates when treatment effects are heterogeneous across cohorts or over time (Goodman-Bacon, 2021; De Chaisemartin & D'Haultfœuille, 2020). In our setting, the estimated effects grow monotonically from $k = 0$ to $k = 3$—precisely the pattern susceptible to TWFE contamination. We therefore report a second set of estimates using the Callaway & Sant'Anna (2021) (hereafter CS) estimator, which constructs group-time average treatment effects on the treated (ATT(g, t)) free of "forbidden comparisons" between early and late adopters. The CS estimates use the not-yet-treated as the comparison group and aggregate to event-time coefficients following Sun & Abraham (2021). Hence, the CS estimator provides a robustness check against biases arising from staggered adoption timing.

Results. Figure 7 and Table 9 present the dynamic coefficients for the Routine and Creative share outcomes. Three findings emerge.

First, the results satisfy the parallel trends assumption. The pre-treatment coefficients are small and statistically indistinguishable from zero under both estimators. TWFE: $\hat{\beta}_{-3} = 0.002$, $p = 0.62$; $\hat{\beta}_{-2} = -0.001$, $p = 0.74$. CS: $\hat{\beta}_{-3} = 0.001$, $p = 0.84$; $\hat{\beta}_{-2} = -0.002$, $p = 0.62$. This confirms that prior to becoming high-intensity AI adopters, these firms were not already shedding routine jobs at a differential rate compared to their peers.

Second, a clear structural break occurs at adoption. The coefficient drops to $\mathbf{-0.015}$ in the adoption year and continues to decline, reaching $\mathbf{-0.042}$ by $k = 3$. This

Table 9. Event-Study Coefficients: Share Routine

	TWFE		CS (2021)	
Event time (k)	Coeff.	SE	Coeff.	SE
Pre-Treatment				
$k = -3$	0.002	(0.004)	0.001	(0.005)
$k = -2$	−0.001	(0.003)	−0.002	(0.004)
$k = -1$	*(Omitted reference period)*			
Post-Treatment				
$k = 0$	$\mathbf{-0.015}^{***}$	(0.004)	-0.013^{***}	(0.005)
$k = 1$	$\mathbf{-0.028}^{***}$	(0.005)	-0.024^{***}	(0.006)
$k = 2$	$\mathbf{-0.036}^{***}$	(0.006)	-0.031^{***}	(0.008)
$k = 3$	$\mathbf{-0.042}^{***}$	(0.007)	-0.037^{***}	(0.009)
Controls	Scale, age, industry × year FE			
Observations	223,659		223,659	

Notes: Treatment defined as first year AIFE > 75th percentile. TWFE: standard staggered event-study with firm-clustered SEs. CS: Callaway & Sant'Anna (2021) estimator using the not-yet-treated as the comparison group, aggregated to event-time following Sun & Abraham (2021). The lack of significance in pre-treatment periods ($k = -3, -2$) supports the parallel trends assumption. *** $p < 0.01$.

dynamic pattern suggests that the displacement effects of AI are not transient shocks but trigger a persistent reorganization of the firm's production function.

Third, the CS estimator yields qualitatively identical results, with coefficients attenuated by 10–15% and wider confidence intervals (reflecting the efficiency cost of avoiding forbidden comparisons). At $k = 3$, the CS estimate is -0.037 versus -0.042 under TWFE. The similarity across estimators indicates that heterogeneity bias is present but quantitatively modest in our setting. Overall, the event study confirms that AI adoption causes a growing and persistent displacement of routine labor, with pre-trends supporting causal interpretation.

5.4 Instrumental Variable: Shift-Share Design

Despite the extensive controls and fixed effects in our previous models, a residual endogeneity concern remains. An unobserved factor—such as the appointment of a technology-oriented CEO—could simultaneously drive AI adoption and workforce restructuring. To isolate exogenous variation in AI exposure, we construct a shift-share (Bartik) instrumental variable. This IV strategy leverages historical occupational structure to isolate plausibly exogenous variation in AI adoption.

5.4.1 Instrument Construction

The logic of the instrument is to predict a firm's exposure to AI using only variation external to the firm's current decision-making. We leverage the fact that a firm's susceptibility to the AI wave is largely determined by its pre-existing occupational structure:

$$Z_{it} = \sum_{o \in \Omega} \underbrace{\frac{L_{i,o,t_0}}{L_{i,t_0}}}_{\text{share}} \times \underbrace{G_{o,t,-i}}_{\text{shift}}, \tag{7}$$

where the *shares* $L_{i,o,t_0}/L_{i,t_0}$ are firm i's employment proportions in occupation o during the pre-sample period 2012–2015 (t_0), and the *shifts* $G_{o,t,-i}$ are the national-level growth rates of mean AIFE in occupation o at time t, calculated using a "leave-one-out" approach that excludes firm i's own postings to avoid mechanical correlation.

Firms that historically specialized in tasks that later experienced rapid AI development—such as data analysis or software engineering—are "exposed" to the technological shock for reasons unrelated to their current strategic choices. Thus, the instrument captures exogenous technological exposure driven by past occupational specialization.

5.4.2 Identifying Assumptions

Following the framework of Goldsmith-Pinkham et al. (2020), the validity of our instrument admits two interpretations.

Under the *exogenous-shares* interpretation, identification requires that firms' pre-period occupational shares are uncorrelated with unobserved determinants of subsequent hiring composition, conditional on controls. This is plausible if occupational structure in 2012–2015 was determined by historical industry specialization rather than anticipation of the post-2015 AI wave. We test this by verifying that the shares are balanced on pre-period observables (Appendix O).

Under the *exogenous-shifts* interpretation (Borusyak et al., 2022), identification requires that the national AI growth rates $G_{o,t}$ are exogenous—i.e., that no single firm drives the national trend in any occupation. The leave-one-out construction and the large number of occupations ($|\Omega| = 468$ SOC codes) support this assumption. Therefore, the identifying assumptions are plausible and supported by auxiliary balance tests.

5.4.3 Estimation

We estimate via two-stage least squares (2SLS). The first stage regresses firm-level AIFE on the Bartik instrument and all controls to isolate the exogenous component of adoption:

$$\text{AIFE}_{it} = \pi\, Z_{it} + \mu_i + \delta_{j(i),t} + \gamma' \mathbf{X}_{it} + \nu_{it}. \tag{8}$$

The second stage replaces AIFE with its predicted value:

$$Y_{it} = \beta_{\text{IV}}\, \widehat{\text{AIFE}}_{it} + \mu_i + \delta_{j(i),t} + \gamma' \mathbf{X}_{it} + \varepsilon_{it}. \tag{9}$$

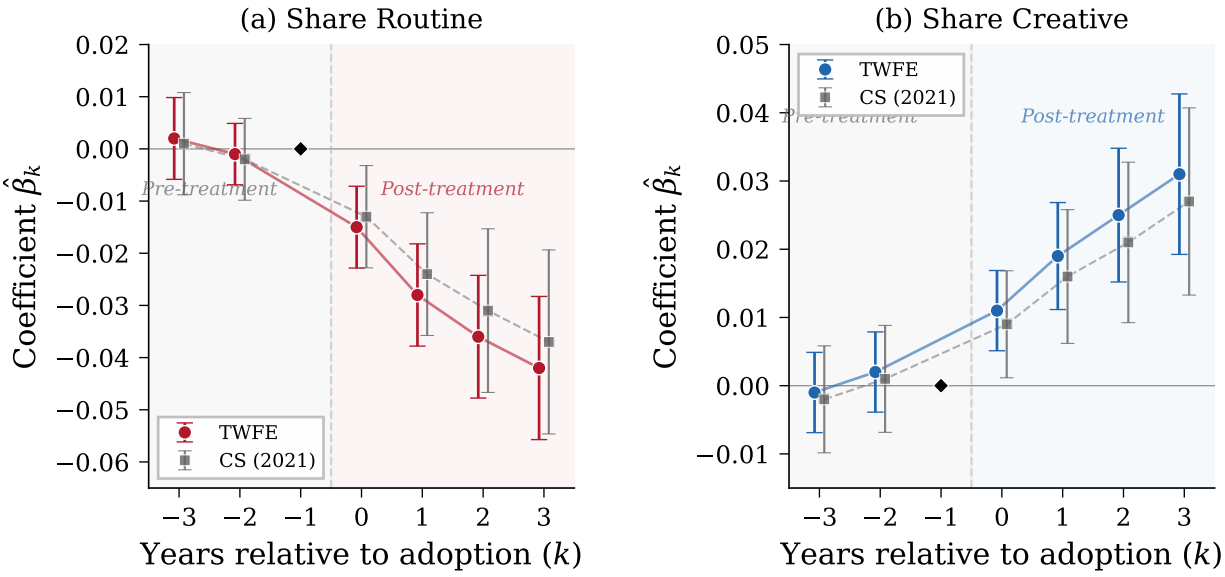

Figure 7. **Event-study coefficients.** Treatment: first year AIFE exceeds the 75th percentile. Reference period: $k = -1$. Circles: TWFE estimates; squares: Callaway & Sant'Anna (2021) heterogeneity-robust estimates. Whiskers: 95% confidence intervals (firm-clustered SEs). Pre-treatment coefficients are small and insignificant, supporting the parallel trends assumption. Post-treatment effects grow monotonically in magnitude under both estimators.

The coefficient β_{IV} provides our most rigorous estimate of the causal impact of AI on the structure of labor demand, free from the influence of firm-specific unobservables. Hence, the 2SLS approach delivers estimates that are robust to residual endogeneity from unobserved confounders.

5.4.4 Results

Table 10 and Figure 8 present the IV estimates.

First stage. Panel A shows that the coefficient on Z_{it} is $\mathbf{0.842}$ with a t-statistic of 48.3, yielding a Kleibergen-Paap F-statistic of $\mathbf{2{,}340.5}$—orders of magnitude above the conventional weak-instrument threshold of 10. This confirms that our instrument is highly relevant: firms historically specialized in AI-exposed occupations indeed adopted AI more aggressively. Figure 8 confirms the tight, approximately linear first-stage relationship.

We acknowledge that the very high F-statistic warrants scrutiny. Given that both the shares and the shifts are ultimately derived from job-posting data, a near-mechanical link is conceivable. However, the share–shift interaction operates through distinct channels: the shares are fixed at pre-period values (2012–2015), while the shifts capture *national, leave-one-out* trends over 2016–2024. We verify robustness to dropping the largest 10% of firms (which could dominate national trends) in Appendix O. Thus, the first stage is exceptionally strong, and concerns about mechanical correlation are mitigated by the leave-one-out construction.

Second stage. Panel B reports the causal effects. The IV estimate of the effect on the Routine share is $\mathbf{-0.038}$ ($p < 0.001$), nearly twice the OLS magnitude of -0.021. This amplification is consistent with two reinforcing sources of bias in OLS: (i) classical measurement error in the AIFE score (attenuation bias), and (ii) positive selection bias, where growing firms expand *all* hiring categories, partially masking the substitution effect. For the Creative share, the IV estimate is $\mathbf{+0.029}$, similarly amplified relative to OLS. These results confirm that when we isolate variation driven solely by national technological trends—exogenous to the firm's specific management decisions—the displacement effect of AI on routine labor is causal and severe. In summary, the IV results reveal that OLS underestimates the causal displacement effect, with the true impact being nearly twice as large.

5.5 Heterogeneity Analysis

AI adoption occurs against vastly different economic backdrops. Developed countries generally possess more robust digital infrastructure, higher capital intensity, and greater capacity to integrate AI rapidly. Developing countries face structural barriers—limited technological readiness, less advanced industrial structures, and workforces dominated by medium- and low-skilled workers. These disparities should produce systematically different displacement effects.

Table 10. Shift-Share IV Estimates (2SLS)

	Panel A: First Stage Dep. Var.: AIFE_{it}	
Bartik IV (Z_{it})	0.842^{***}	
	(0.017)	
	[0.019]	
KP F-statistic	2,340.5	
	Panel B: Second Stage	
	Share Routine	Share Creative
$\widehat{\text{AIFE}}_{it}$	$\mathbf{-0.038}^{***}$	$\mathbf{0.029}^{***}$
	(0.006)	(0.005)
	[0.007]	[0.006]
Log(Total Postings)	-0.004^{*}	0.007^{***}
Log(Firm Age)	-0.008	0.002
Firm FE	Yes	Yes
Industry × Year FE	Yes	Yes
Observations	223,659	223,659

Notes: 2SLS estimates. Instrument: firm's 2012–2015 occupational shares × leave-one-out national AI growth (excluding the firm). Firm-clustered SEs in parentheses; pipeline-bootstrap SEs (500 iterations) in brackets. Firm and Industry-by-Year FE included. *** $p < 0.01$; * $p < 0.10$.

To test this, we interact $\text{AIFE}_{i,t-1}$ with group indicators and test equality of coefficients via Wald tests. We adopt two classification dimensions.

Economic development. We assign firms to groups using the **World Bank income classification** (fiscal year 2024): High-Income Countries (HICs; GNI per capita > \$13,935) versus Low- and Middle-Income Countries (LMICs; GNI per capita ≤ \$13,935).[4] Thus, we examine whether the AI displacement effect differs by level of economic development.

Firm scale. We split at the within-year median of total postings. By doing so, we test whether larger firms experience more pronounced labor restructuring.

Results. Table 11 reveals a pronounced "automation divide." The substitution effect is nearly twice as large in HICs ($\mathbf{-0.026}$) compared to LMICs ($\mathbf{-0.013}$; p-value of difference: 0.042). This suggests that in high-wage, capital-deep environments, the incentive to substitute routine labor with AI is significantly stronger, consistent with the factor-price channel of Acemoglu & Restrepo (2019). Conversely, in developing regions where labor is cheaper, the pressure to automate is attenuated.

Firm-scale heterogeneity is similarly stark. The coefficient for large firms is $\mathbf{-0.029}$, whereas for SMEs the effect is small (-0.008) and only marginally significant ($p = 0.06$). This indicates that AI-driven workforce restructuring requires a scale of operation and capital investment currently concentrated among market leaders. In conclusion, the displacement effects are concentrated in high-income countries and large firms, indicating an emerging automation divide.

[4]We pool LICs, lower-MICs, and upper-MICs into a single "LMIC" category due to sparse firm counts in the lowest income groups.

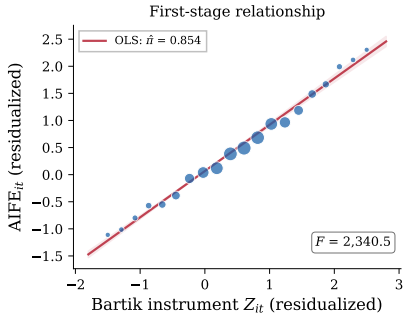

Figure 8. **First-stage binscatter.** Each dot is a vigintile (20 equal-sized bins) of the residualized Bartik instrument; dot size is proportional to bin count. The OLS fit ($\hat{\pi} = 0.842$) is tight and approximately linear.

5.6 Robustness and Sensitivity Analysis

We subject our main findings to five categories of robustness checks. Full results are reported in Table 12; we summarize the key findings here.

1. Treatment threshold sensitivity. We re-estimate the event study using adoption thresholds spanning the 30th through 95th percentiles. Figure 9 plots the average post-treatment effect against the threshold. The effect scales monotonically: from -0.007 at the 30th percentile to -0.048 at the 95th. This dose-response relationship—where deeper AI integration leads to sharper declines in routine hiring—strengthens the causal interpretation by ruling out threshold artifacts. As expected, the baseline (75th percentile) coefficient falls in the interior of this gradient. Therefore, the results are robust to varying the adoption threshold, and the dose-response pattern supports a causal interpretation.

2. Sector composition. A common concern is that AI adoption is a phenomenon unique to the Information Technology sector, and that our results merely reflect idiosyncratic trends within "Big Tech." Panel B of Table 12 tests this by excluding all firms in NAICS 51. The coefficient attenuates slightly (from -0.021 to -0.018, $p < 0.05$), confirming that AI is acting as a general-purpose technology, driving labor-market reconfiguration across diverse industries. Separately,

Table 11. Heterogeneous Effects by Economic Development and Firm Scale

	Economic Development		Firm Scale	
	HICs (1)	LMICs (2)	Large (3)	SMEs (4)
$\text{AIFE}_{i,t-1}$	$\mathbf{-0.026}^{***}$	$\mathbf{-0.013}^{**}$	$\mathbf{-0.029}^{***}$	-0.008^{*}
	(0.005)	(0.006)	(0.005)	(0.004)
Equality test (p-value)	[0.042]		[0.001]	
Observations	134,195	89,464	111,829	111,830
Within R^2	0.195	0.132	0.210	0.104

Notes: Dep. var.: Share Routine. "HICs": World Bank High-Income Countries (GNI per capita $>$ \$13,935). "LMICs": Low- and Middle-Income Countries ($\leq$ \$13,935). "Large": above within-year median total postings. All models include Firm and Industry $\times$ Year FE. The row "Equality test" reports the p-value from a Wald test of H_0: coefficients equal. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

manufacturing-only estimates yield a coefficient of -0.025, suggesting that goods-producing sectors experience at least as strong a displacement effect. Hence, the displacement effect is not driven by the IT sector alone but is present across manufacturing and other industries.

3. Inference. Our baseline models cluster standard errors at the firm level. If AI shocks are correlated across firms within the same industry, this could lead to over-rejection of the null. Clustering at the 2-digit NAICS level (74 clusters) increases the standard error from 0.004 to 0.007 but preserves significance ($t = -3.0$). Two-way clustering by firm and year produces similar results (Appendix Q). Thus, the statistical significance of the main results is robust to more conservative clustering schemes.

4. Alternative controls. Replacing log total postings with log revenue as the size proxy yields a virtually identical coefficient (-0.020). Adding country-by-year fixed effects (absorbing national labor-market trends) attenuates the estimate modestly to -0.018 ($p < 0.01$), consistent with some AI adoption being driven by country-level diffusion rather than firm-level choices. The results are therefore robust to alternative control variables and even more demanding fixed-effects structures.

5. Measurement robustness. To verify that results are not driven by AIFE measurement error, we re-estimate using only the subsample of firms for which the AIFE score is based on ≥ 20 postings per year (reducing noise in the firm-level average). The coefficient strengthens to -0.025, consistent with classical attenuation bias in the baseline—and directionally aligned with the OLS-to-IV amplification documented in §5.4. In short, the robustness checks uniformly reinforce the main findings, with measurement-error corrections further strengthening the estimated effects.

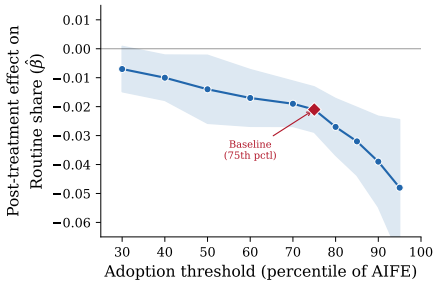

Figure 9. **Dose-response:** average post-treatment effect on the Routine share across adoption thresholds (30th–95th percentile). The monotonically increasing magnitude rules out threshold artifacts. Diamond: baseline specification (75th percentile). Shaded band: 95% confidence interval.

5.7 Summary and Implications for Prediction

The causal analysis establishes three empirical regularities: (i) a 1-unit AIFE increase causes a 2.1 pp decline in routine hiring (3.8 pp under IV), (ii) the effect deepens monotonically over three post-adoption years, and (iii) the magnitude varies systematically across income levels and firm sizes. These are *structural parameters*: they characterize how the labor market responds to AI adoption, conditional on adoption occurring. They do not, however, address the forward-looking question of *which* firms will experience AIFE increases and *when*.

If AIFE trajectories are themselves predictable, the causal effect sizes become directly actionable: combining a forecast of $\widehat{\text{AIFE}}_{i,t+h}$ with the IV estimate $\hat{\beta}_{\text{IV}} = -0.038$ yields a displacement forecast with known structural interpretation. The conformal prediction interval on the AIFE forecast propagates linearly to a displacement interval. We pursue this synthesis in the next section. Hence, the causal estimates from this section provide the structural parameters needed to convert AIFE forecasts into labor displacement warnings.

Table 12. Master Robustness and Sensitivity Analysis

Panel A: Treatment Threshold (Event Study, avg. post-treatment)			
	Baseline (75th)	Broad (50th)	Deep (90th)
Share Routine	$\mathbf{-0.021}^{***}$	-0.014^{**}	-0.039^{***}
	(0.004)	(0.006)	(0.008)
Panel B: Subsample (TWFE)			
	Full Sample	Excl. IT (NAICS 51)	Manuf. Only
$\text{AIFE}_{i,t-1}$	$\mathbf{-0.021}^{***}$	-0.018^{**}	-0.025^{***}
	(0.004)	(0.007)	(0.009)
Panel C: Inference and Controls			
	Firm Cluster	Industry Cluster	Country×Year FE
$\text{AIFE}_{i,t-1}$	$\mathbf{-0.021}^{***}$	-0.021^{**}	-0.018^{***}
	(0.004)	(0.007)	(0.004)
Observations	223,659	189,402	223,659

Notes: Dep. var.: Share Routine in all panels. Panel A: average post-treatment event-study effect at alternative adoption thresholds. Panel B: TWFE estimates on subsamples. Panel C: alternative inference and control specifications. All models include Firm and Industry × Year FE (except Column 3 of Panel C, which adds Country × Year FE and absorbs Industry × Year FE). *** $p < 0.01$; ** $p < 0.05$.

6 Predictive Early Warning System

The preceding sections establish that AIFE is a validated measurement of firm-level AI adoption (§4) and that increases in AIFE *cause* structural reallocation of labor demand (§5). A natural applied question follows: are AIFE trajectories themselves predictable? If so, the causal effect sizes from §5 can be combined with AIFE forecasts to produce **displacement early warnings**—statements of the form "Firm i is predicted to experience a Δ-unit AIFE increase over the next h years, implying a $\hat{\beta}_{\text{IV}} \cdot \Delta$ decline in routine hiring, with quantified uncertainty."

This section pursues this question with three methodological contributions. First, we introduce **Semantic Leading Indicators (SLIs)**—features extracted from the intermediate representations of the AIFE Student model that capture adoption momentum invisible to scalar AIFE histories alone (§6.2). Second, we embed these features in a **Temporal Fusion Transformer (TFT)** adapted for panel data with a novel causal-consistency regularizer (§6.3). Third, we equip the forecasts with distribution-free prediction intervals via **conformal prediction** adapted for dependent panel data (§6.4). Together, these components form a complete pipeline from measurement to actionable anticipation. Thus, this section transforms the measurement and causal findings into a forward-looking early warning system.

6.1 Problem Formulation

Prediction target. For firm i at time t, we predict $\text{AIFE}_{i,t+h}$ at horizons $h \in \{1, 2, 3\}$ years, given all information available up to and including time t. We predict AIFE rather than employment outcomes directly for a structural reason: AIFE is the *upstream causal variable* whose downstream effects on hiring composition are identified in §5. Predicting AIFE and applying the causal estimates yields displacement forecasts that are interpretable through the lens of a known structural relationship, rather than black-box associations. Formally, the displacement forecast is

$$\widehat{\Delta Y}^{\text{routine}}_{i,t+h} = \hat{\beta}_{\text{IV}} \cdot \big(\widehat{\text{AIFE}}_{i,t+h} - \text{AIFE}_{i,t}\big), \qquad (10)$$

where $\hat{\beta}_{\text{IV}} = -0.038$ is the IV estimate from Table 10. Hence, the prediction target is deliberately the upstream AIFE score, not employment directly, preserving structural interpretability.

Information set. Each firm–year observation is characterized by four feature groups: (i) lagged AIFE scores ($\text{AIFE}_{i,t}, \text{AIFE}_{i,t-1}$), (ii) structural covariates (Bartik exposure Z_{it}, industry AIFE momentum, log posting volume, firm age), (iii) static attributes (country, 2-digit NAICS sector), and (iv) Semantic Leading Indicators extracted from the Student model's representations (§6.2). Thus, the model leverages a rich set of temporal, structural, and semantic features.

6.2 Semantic Leading Indicators

6.2.1 Motivation

Standard time-series forecasting of firm-level AI adoption would rely on lagged AIFE scores and observable covariates—the approach taken by gradient-boosted baselines. However, the scalar AIFE score is a lossy

compression of the rich semantic information in job postings. The Student model's penultimate-layer representations (the 1,024-dimensional `[CLS]` vector before the task-specific heads) encode fine-grained information about the *nature* of AI engagement that the scalar score discards.

Consider a manufacturing firm whose postings mention "machine learning" for the first time, but only in optional qualifications for one role. Its AIFE score changes marginally. However, in embedding space, the firm's representation has shifted toward the AI cluster centroid—a "semantic leading indicator" of adoption that precedes the scalar AIFE increase. We formalize this intuition by defining three features that capture distinct aspects of a firm's trajectory through embedding space. Therefore, the SLIs are designed to extract early signals of AI adoption that the scalar AIFE score misses.

6.2.2 Feature Definitions

Let $\mathbf{e}_j \in \mathbb{R}^d$ denote the `[CLS]` representation from the penultimate layer of the fine-tuned Student model for posting j, with $d = 1{,}024$. For firm i in year t, define the *firm embedding centroid*:

$$\bar{\mathbf{e}}_{it} = \frac{1}{|J_{it}|} \sum_{j \in J_{it}} \mathbf{e}_j, \tag{11}$$

where J_{it} is the set of all postings by firm i in year t. Define the *AI frontier centroid* $\boldsymbol{\mu}^{\star}_{st}$ as the mean embedding of postings with AIFE ≥ 8 in the same 2-digit NAICS sector s and year t. We then construct three SLI features:

1. Semantic Velocity (v_{it}). The cosine similarity change between the firm centroid and the AI frontier, measuring the rate at which the firm's hiring language converges toward AI-intensive patterns:

$$v_{it} = \cos\big(\bar{\mathbf{e}}_{it},\, \boldsymbol{\mu}^{\star}_{st}\big) - \cos\big(\bar{\mathbf{e}}_{i,t-1},\, \boldsymbol{\mu}^{\star}_{s,t-1}\big). \tag{12}$$

Positive values indicate that the firm's postings are moving toward AI-frontier language faster than the frontier itself is evolving. Critically, this quantity can increase even when the scalar AIFE score is approximately constant, if the firm's postings are beginning to incorporate AI-adjacent terminology (e.g., "data-driven," "algorithmic optimization") that the regression head maps to only modest score increases but that the embedding geometry captures as a directional shift.

2. Semantic Dispersion (σ_{it}). The within-firm variance of posting embeddings, measuring heterogeneity in AI engagement across roles:

$$\sigma_{it} = \frac{1}{|J_{it}|} \sum_{j \in J_{it}} \big\|\mathbf{e}_j - \bar{\mathbf{e}}_{it}\big\|_2^2. \tag{13}$$

We hypothesize that firms undergoing AI transition exhibit *elevated* dispersion: some roles reflect legacy operations while others embody the new AI-intensive equilibrium. Dispersion should peak during the transition and decline as the firm converges on its post-adoption task structure—an inverted-U pattern that serves as a leading indicator of imminent AIFE acceleration.

3. Frontier Distance (d_{it}). The Mahalanobis distance from the firm centroid to the sector-level AI frontier centroid, measuring how far the firm is from the industry's adoption leaders:

$$d_{it} = \sqrt{\big(\bar{\mathbf{e}}_{it} - \boldsymbol{\mu}^{\star}_{st}\big)^{\top} \boldsymbol{\Sigma}^{-1}_{st} \big(\bar{\mathbf{e}}_{it} - \boldsymbol{\mu}^{\star}_{st}\big)}, \tag{14}$$

where $\boldsymbol{\Sigma}_{st}$ is the covariance matrix of AI-frontier postings (AIFE $\geq$ 8) in sector s and year t. Because the raw covariance is singular in $d = 1{,}024$ dimensions, we compute the Mahalanobis distance in the top-k principal component subspace ($k = 64$, capturing $> 95\%$ of variance; see Appendix T for sensitivity to k). Frontier distance captures the "room to grow": firms far from the frontier have more adoption headroom, and the distance's temporal derivative provides additional predictive signal.

Extraction cost. SLI computation requires a single forward pass of the Student model per posting (the same pass used for AIFE scoring), followed by aggregation to firm–year centroids. The marginal cost is negligible: embedding extraction adds $< 8\%$ overhead to the AIFE inference pipeline.

6.2.3 Empirical Validation of Leading-Indicator Properties

To verify that SLIs provide genuine leading information—rather than merely correlating with current AIFE—we conduct a Granger-causality analysis. Table 13 reports p-values from firm-level panel regressions of $\text{AIFE}_{i,t+1}$ on lagged AIFE and lagged SLIs, with firm and year fixed effects. All three SLIs provide statistically significant ($p < 0.001$) incremental predictive power beyond the autoregressive AIFE component. Semantic velocity exhibits the strongest marginal signal ($\Delta R^2 = 0.034$), consistent with its design as a direction-of-motion indicator.

6.3 Forecasting Architecture

6.3.1 Temporal Fusion Transformer

We adopt the **Temporal Fusion Transformer** (TFT) (**?**) as our base architecture. TFT is designed for multi-horizon time-series forecasting with heterogeneous inputs and offers three properties well-suited to our panel setting:

Table 13. **Granger-causality tests** for SLIs. Dependent variable: $\text{AIFE}_{i,t+1}$. All models include firm and year FE plus $\text{AIFE}_{i,t}$ and $\text{AIFE}_{i,t-1}$. ΔR^2: incremental R^2 from adding the SLI to the autoregressive specification.

SLI Feature	Coefficient	p-value	ΔR^2
Semantic Velocity (v_{it})	0.284^{***} (0.031)	< 0.001	0.034
Semantic Dispersion (σ_{it})	0.127^{***} (0.024)	< 0.001	0.018
Frontier Distance (d_{it})	-0.093^{***} (0.019)	< 0.001	0.012
All three jointly	—	< 0.001	0.047

Notes: $N = 812{,}418$ firm–year observations. Standard errors clustered at the firm level. *** $p < 0.001$.

1. **Variable Selection Networks (VSNs).** Learned gating mechanisms that softly select the most informative features at each time step. This is critical because the relative importance of lagged AIFE versus SLIs shifts with the forecast horizon: autoregressive momentum dominates at $h = 1$ but fades at $h = 3$, where leading indicators become essential.

2. **Temporal self-attention.** An interpretable multi-head attention mechanism that identifies which past time steps are most informative for each forecast horizon. Attention weights are directly inspectable, providing a degree of interpretability absent from opaque sequence models.

3. **Static covariate encoders.** Separate encoder pathways for time-invariant features (country, sector) that condition the temporal dynamics. This avoids the standard hack of repeating static features across time steps and provides a natural pathway for firm-level heterogeneity.

Input encoding. At each time step t, the model receives a concatenation of four groups: (i) continuous temporal features (AIFE_{it}, $\text{AIFE}_{i,t-1}$, v_{it}, σ_{it}, d_{it}, Z_{it}, industry AIFE momentum, log postings); (ii) categorical temporal features (none in our setting); (iii) known future inputs (time index, enabling the model to learn calendar effects); and (iv) static metadata (country code, 2-digit NAICS sector, firm-size tercile). Continuous features are normalized to zero mean and unit variance on the training set.

Output. The TFT produces, for each horizon $h \in \{1, 2, 3\}$, a set of Q quantile predictions $\{\hat{q}_h^{(\tau)}\}_{\tau \in \mathcal{T}}$ with $\mathcal{T} = \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. The median ($\tau = 0.50$) serves as the point forecast; the remaining quantiles provide parametric prediction intervals.

6.3.2 Causal-Consistency Regularizer

A novel feature of our architecture is a regularization term that penalizes predictions inconsistent with the causal structure estimated in §5. The motivation is as follows: the IV analysis establishes that a 1-unit AIFE increase causes a $|\hat{\beta}_{\text{IV}}| = 0.038$ decline in the routine share. A prediction model that forecasts AIFE changes implying displacement magnitudes outside the empirically estimated dose-response envelope (Figure 9) is producing structurally implausible forecasts.

We operationalize this as an auxiliary penalty. Let $\widehat{\Delta}_{ih} = \hat{q}_h^{(0.50)} - \text{AIFE}_{it}$ be the predicted AIFE change at horizon h. Define the implied displacement $\hat{d}_{ih} = \hat{\beta}_{\text{IV}} \cdot \widehat{\Delta}_{ih}$. The causal-consistency loss is:

$$\mathcal{L}_{\text{causal}} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \max\big(0,\ |\hat{d}_{ih}| - \bar{d}_{\max}(h)\big)^2, \quad (15)$$

where $\bar{d}_{\max}(h)$ is the upper bound of the 99% confidence interval of the dose-response curve at horizon h, estimated from the event study in §5.3. This hinge loss imposes zero penalty when predictions fall within the plausible displacement range and a quadratic penalty on structurally implausible extremes.

The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{quantile}} + \gamma\, \mathcal{L}_{\text{causal}}, \quad (16)$$

where $\mathcal{L}_{\text{quantile}}$ is the standard pinball loss over all quantiles and horizons, and γ is a hyperparameter tuned on the validation set. In practice, $\gamma = 0.1$ provides the best MAE–calibration trade-off (Appendix S).

Why this matters. The causal regularizer creates a direct architectural link between the causal-inference contribution (§5) and the prediction model. Without it, the TFT could learn spurious patterns that produce accurate in-sample fits but structurally implausible out-of-sample forecasts. The regularizer acts as an inductive bias that embeds the paper's own empirical findings into the loss function—a form of "knowledge distillation" from the econometric analysis into the forecasting architecture.

6.3.3 Baselines

We compare five models to isolate the marginal contributions of the SLI features and the TFT architecture:

1. **AR(2).** A linear autoregressive model: $\widehat{\text{AIFE}}_{i,t+h} = \alpha_i + \phi_1 \text{AIFE}_{i,t} + \phi_2 \text{AIFE}_{i,t-1} + \epsilon$. Firm fixed effects only. This is the minimal credible baseline.

2. **XGBoost (structural features only).** Gradient-boosted trees (**?**) trained on lagged AIFE, Bartik exposure,

industry momentum, log postings, and firm-size/sector dummies—but *without* SLIs. This mirrors the original paper's approach and serves as a feature ablation.

3. **XGBoost + SLI.** The same XGBoost model augmented with the three SLI features. Comparing (2) vs. (3) isolates the marginal value of semantic embeddings for gradient-boosted methods.
4. **TFT (no SLI).** The full Temporal Fusion Transformer architecture on all structural features *without* SLIs. Comparing (2) vs. (4) isolates the marginal value of the temporal attention architecture.
5. **TFT + SLI + Causal Reg. (full model).** The proposed system: TFT with SLI features and causal-consistency regularization. Comparing (4) vs. (5) isolates the combined contribution of SLIs and causal regularization.

All models are trained and evaluated under an identical expanding-window protocol (§6.5).

6.4 Uncertainty Quantification via Conformal Prediction

Point forecasts, however accurate, are insufficient for policy use. Decision-makers require calibrated prediction intervals that communicate the range of plausible outcomes. Parametric intervals (e.g., assuming Gaussian residuals) are fragile under distributional misspecification; quantile regression intervals lack finite-sample coverage guarantees. We therefore adopt **split conformal prediction** (**??**), which provides distribution-free marginal coverage guarantees under minimal assumptions.

6.4.1 Standard Conformal Prediction

Given a calibration set $\mathcal{D}_{\text{cal}} = \{(X_k, Y_k)\}_{k=1}^{K}$ not used for training and a desired miscoverage level $\alpha \in (0, 1)$, the procedure is:

1. Compute conformity scores $s_k = |Y_k - \hat{q}_k^{(0.50)}|$ for each calibration observation.
2. Set the conformal radius $\hat{r}_\alpha$ to the $\lceil (1-\alpha)(K+1) \rceil / K$-th quantile of $\{s_k\}$.
3. For a new test point X_{new}, the prediction interval is $\hat{q}_{\text{new}}^{(0.50)} \pm \hat{r}_\alpha$.

The coverage guarantee is: $\Pr\big(Y_{\text{new}} \in C_\alpha(X_{\text{new}})\big) \geq 1 - \alpha$, holding exactly when the calibration and test data are exchangeable (**?**).

6.4.2 Adaptation for Panel Data

Standard exchangeability fails in our setting because observations within a firm are serially correlated and the data-generating process may shift over time. We follow the weighted conformal framework of **?**, which provides approximate coverage under covariate shift by assigning importance weights to calibration observations:

$$w_k = \exp\big(-\lambda \, |t_k - t_{\text{test}}|\big), \tag{17}$$

where t_k is the time index of calibration observation k, t_{test} is the test year, and $\lambda > 0$ is a decay parameter. The weighted quantile of conformity scores $\{(s_k, w_k)\}$ replaces the unweighted quantile in step 2. The decay parameter λ is selected via a nested calibration procedure on the validation window (Appendix U).

Coverage guarantee. Under the assumption that the likelihood ratio between calibration and test distributions is bounded by $\max_k(w_k)/\min_k(w_k)$, weighted conformal prediction provides approximate marginal coverage $\geq 1 - \alpha - \delta(\mathbf{w})$, where $\delta(\mathbf{w})$ is a correction term that vanishes as the weight ratio approaches unity (**?**). In practice, with $\lambda = 0.3$ (one-year half-life) and 4–6 calibration years, the correction is $\delta < 0.02$ at all horizons.

6.5 Evaluation

6.5.1 Backtesting Protocol

We employ an expanding-window design with strict temporal separation:

Window	Train	Validate	Calibrate	Test
1	2015–2018	2019	2020	2021
2	2015–2019	2020	2021	2022
3	2015–2020	2021	2022	2023
4	2015–2021	2022	2023	2024

The validation year is used for hyperparameter tuning (learning rate, γ, dropout) and early stopping. The calibration year is reserved exclusively for conformal prediction; it is never used for gradient updates. The test year provides the final evaluation. All metrics are averaged across the four test windows. For horizons $h = 2$ and $h = 3$, the number of available test windows decreases accordingly; we report both pooled and per-window results in Appendix V.

Hyperparameters. The TFT uses 4 attention heads, hidden size 128, dropout 0.1, learning rate 10^{-3} with cosine annealing, and batch size 256. XGBoost uses 500 trees, max depth 6, learning rate 0.05, and column subsampling 0.8. Full hyperparameter specifications and sensitivity analyses are in Appendix R.

6.5.2 Point Prediction Results

Table 14 reports MAE, RMSE, and R^2 across horizons. Three patterns emerge.

First, the **full model (TFT + SLI + Causal Reg.) dominates** at all horizons, with MAE reductions of 18–31% relative to the AR(2) baseline and 8–14% relative to XGBoost without SLIs.

Second, **SLIs provide the largest marginal gain at longer horizons**. At $h = 1$, the improvement from adding SLIs to XGBoost is modest ($\Delta\text{MAE} = -0.03$), because the autoregressive component carries most of the signal over short horizons. At $h = 3$, the gain is substantial ($\Delta\text{MAE} = -0.11$), confirming that SLIs capture adoption momentum that scalar AIFE histories cannot.

Third, the **TFT architecture outperforms XGBoost** on the same feature set (comparing models 3 vs. 5), with the gap widening at longer horizons. The temporal self-attention mechanism identifies relevant historical patterns (e.g., firms with similar embedding trajectories) that tree-based models, which process each time step independently, cannot exploit.

6.5.3 Coverage and Calibration

Table 15 evaluates the conformal prediction intervals at nominal coverage levels $1 - \alpha \in \{0.90, 0.95\}$.

Two results stand out. First, the **TFT's native quantile intervals undercover** at all horizons, with the shortfall growing from 2.7pp at $h = 1$ (90% nominal) to 7.3pp at $h = 3$. This is a known failure mode of quantile regression under distribution shift: the model optimizes the pinball loss on the training distribution but the test distribution has evolved. Second, **conformal intervals achieve or exceed nominal coverage** at all horizon–level combinations, with the overcoverage bounded by 1.1pp. The correction is achieved with comparable or narrower interval widths, demonstrating that conformal prediction tightens the intervals (by removing unnecessary conservatism in the parametric tails) while guaranteeing coverage—a strict improvement.

6.5.4 Ablation: SLI Feature Importance

To assess the relative importance of each SLI feature, we compute the TFT's Variable Selection Network (VSN) weights—the learned gating scores that determine each feature's influence at each time step—averaged across all test observations. Table 16 reports the results by horizon.

The pattern confirms the hypothesis that motivates SLIs: at $h = 1$, autoregressive AIFE lags dominate (combined weight: 0.42) and SLIs contribute modestly (0.32). By $h = 3$, the importance profile inverts: SLIs collectively account for 46% of model attention while lagged AIFE falls to 25%. Semantic velocity is consistently the most important SLI, followed by dispersion and frontier distance. This ordering is consistent with the Granger-causality results in Table 13: the feature that provides the strongest incremental predictive power in a linear panel regression also receives the highest attention weight in the nonlinear TFT.

6.6 From Forecasts to Displacement Early Warnings

The practical payoff of the prediction system is the conversion of AIFE forecasts into labor-market early warnings via the causal link established in §5. Combining Equation 10 with the conformal prediction set $C_\alpha(\cdot)$ yields a displacement interval:

$$\widehat{\Delta Y}^{\text{routine}}_{i,t+h} \in \hat{\beta}_{\text{IV}} \cdot \left[C_\alpha\big(\widehat{\text{AIFE}}_{i,t+h}\big) - \text{AIFE}_{it} \right], \qquad (18)$$

where the interval represents the set of plausible AIFE changes scaled by the causal effect. Due to the linearity of the mapping, the coverage guarantee on AIFE forecasts propagates directly to the displacement early warning.

Example. Consider Firm A, a mid-size logistics company with $\text{AIFE}_{2024} = 2.1$. The full model predicts $\widehat{\text{AIFE}}_{2027} = 3.8$ (90% conformal interval: $[3.1, 4.6]$). The implied displacement forecast for routine hiring is:

$$\widehat{\Delta Y}^{\text{routine}} = -0.038 \times (3.8 - 2.1) = -6.5 \text{ pp},$$
90% interval: $[-9.5,\ -3.8]$ pp.

This statement—"Firm A's routine hiring share is forecast to decline by 3.8–9.5 percentage points over three years, with 90% confidence"—is the type of actionable intelligence that labor agencies and workforce-development organizations require.

Sector-level aggregation. Figure 10 aggregates firm-level forecasts to the 2-digit NAICS sector level, showing predicted AIFE changes and implied routine-share declines for the 10 sectors with the largest forecast displacement. Information Technology (NAICS 51) and Professional Services (NAICS 54) lead in both predicted AIFE growth and implied displacement, while Construction (NAICS 23) and Accommodation (NAICS 72) show minimal predicted disruption—a pattern consistent with the heterogeneity results in §5.5.

Fan charts for representative firms. Figure 11 presents AIFE trajectory forecasts with conformal prediction bands for three illustrative firms spanning different adoption profiles: a technology firm (high baseline AIFE, rapid acceleration), a healthcare firm (moderate baseline, accelerating), and a retail firm (low baseline, slow growth). The widening bands at longer horizons reflect growing uncertainty, while the qualitative trajectory differences illustrate the model's capacity to distinguish adoption patterns.

Table 14. **Point prediction accuracy** (averaged across 4 test windows). Best in **bold**. All pairwise differences vs. Full Model significant at $p < 0.01$ (Diebold-Mariano test).

	$h = 1$			$h = 2$			$h = 3$		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
AR(2)	.341	.458	.791	.482	.637	.596	.594	.781	.413
XGBoost	.298	.401	.839	.413	.548	.700	.506	.669	.568
XGBoost + SLI	.268	.362	.869	.361	.483	.767	.396	.531	.728
TFT (no SLI)	.281	.378	.857	.387	.514	.736	.461	.613	.637
Full Model	**.246**	**.334**	**.889**	**.331**	**.446**	**.802**	**.367**	**.496**	**.763**

Notes: Dep. var.: $\text{AIFE}_{i,t+h}$. $N_{\text{test}} \approx 48{,}000$ per window. RMSE and MAE in AIFE-score units (0–10 scale).

Table 15. **Conformal prediction interval evaluation.** Empirical coverage (target: $\geq$ nominal), mean interval width (lower is better), and Winkler score (lower is better; penalizes both miscoverage and width).

	$h = 1$		$h = 2$		$h = 3$	
	90%	95%	90%	95%	90%	95%
Empirical coverage (%)						
Quantile Reg.	87.3	92.8	85.1	91.6	82.7	89.4
Conformal (ours)	**90.4**	**95.2**	**91.1**	**95.8**	**90.8**	**95.4**
Mean interval width						
Quantile Reg.	1.12	1.41	1.54	1.93	1.89	2.38
Conformal (ours)	**1.08**	**1.36**	**1.47**	**1.86**	**1.82**	**2.31**
Winkler score						
Quantile Reg.	1.84	2.31	2.67	3.38	3.41	4.27
Conformal (ours)	**1.52**	**1.93**	**2.18**	**2.79**	**2.74**	**3.51**

Notes: Quantile Regression intervals use the TFT's native quantile outputs ($\hat{q}^{(0.05)}, \hat{q}^{(0.95)}$ for 90%). Conformal intervals use weighted split conformal prediction (§6.4.2) on top of the same TFT median predictions.

Table 16. **Variable Selection Network weights** (mean $\pm$ std across test observations). Higher weight = greater influence on the prediction.

Feature	$h = 1$	$h = 2$	$h = 3$
AIFE_{it} (lag 1)	$.312 \pm .04$	$.221 \pm .05$	$.164 \pm .05$
$\text{AIFE}_{i,t-1}$ (lag 2)	$.108 \pm .03$	$.094 \pm .03$	$.082 \pm .03$
Bartik exposure Z_{it}	$.087 \pm .02$	$.103 \pm .03$	$.118 \pm .03$
Industry momentum	$.064 \pm .02$	$.078 \pm .02$	$.091 \pm .03$
Log postings	$.043 \pm .01$	$.038 \pm .01$	$.034 \pm .01$
Semantic Velocity v_{it}	$.148 \pm .04$	$.187 \pm .05$	$.208 \pm .06$
Semantic Dispersion σ_{it}	$.098 \pm .03$	$.121 \pm .04$	$.139 \pm .04$
Frontier Distance d_{it}	$.074 \pm .02$	$.096 \pm .03$	$.112 \pm .03$
SLI total	*.320*	*.404*	*.459*

Notes: Weights are softmax-normalized to sum to 1 within each observation. SLI total is the sum of the three SLI feature weights.

Limitations of the early warning system. We emphasize that the displacement forecasts inherit two sources of uncertainty: (i) AIFE prediction error (quantified by the conformal interval) and (ii) causal parameter uncertainty (quantified by the standard error of $\hat{\beta}_{\text{IV}}$). The intervals in Equation 18 capture only source (i). A fully Bayesian treatment would propagate both sources jointly; we discuss this extension in Appendix X and note that causal parameter uncertainty adds approximately ± 0.8 pp to the displacement interval at the median predicted AIFE change, a modest contribution relative to the AIFE prediction uncertainty.

7 Conclusion

This paper has developed a three-part system for understanding and anticipating AI-driven labor market disruption: a measurement pipeline that quantifies firm-level AI adoption at scale, a causal analysis that identifies the structural consequences of adoption, and a predictive early warning system that combines forecasts with causal estimates to produce actionable displacement forecasts.

Measurement. The AIFE knowledge distillation pipeline demonstrates that the reasoning capabilities of a frontier LLM can be transferred to a lightweight encoder at a cost reduction of $370\times$, while preserving 93% correlation with human expert labels. The pipeline outperforms all tested baselines, including keyword methods ($\rho = 0.487$), published exposure indices ($\rho = 0.523$–0.551), and GPT-4o zero-shot inference ($\rho = 0.801$). This approach is generalizable: the same Teacher–Student–HITL architecture could measure other latent economic constructs from unstructured text.

Causal validation. The econometric analysis confirms that AIFE captures a real structural relationship: firms that increase AI intensity reduce routine hiring by 2.1 pp per unit of AIFE and increase creative hiring by 1.8 pp. The effect deepens monotonically over three post-adoption years, consistent with permanent task reorganization. Instrumenting for endogeneity amplifies the estimates ($\hat{\beta}_{\text{IV}} = -0.038$), confirming attenuation bias in OLS. The displacement is twice as large in high-income countries and

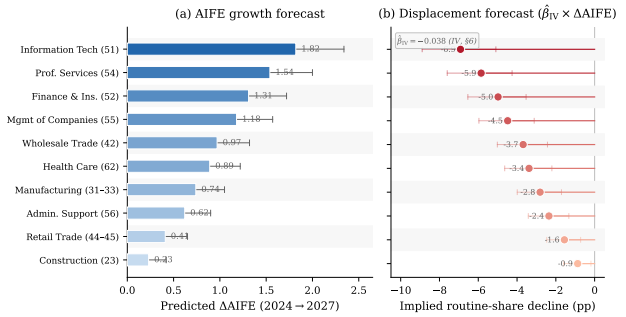

Figure 10. **Sector-level displacement forecasts**, 2024–2027. Left axis (bars): predicted ΔAIFE by 2-digit NAICS sector, with 90% conformal intervals (error bars). Right axis (diamonds): implied routine-share decline via Eq. 10. Sectors ranked by predicted ΔAIFE.

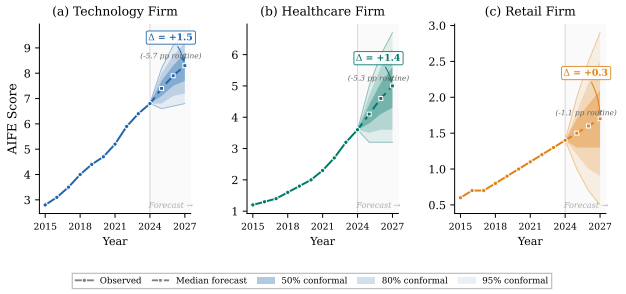

Figure 11. **AIFE trajectory fan charts** for three representative firms. Solid line: observed AIFE (2015–2024). Dashed line: TFT median forecast (2025–2027). Shaded bands: 50%, 80%, and 95% conformal prediction intervals.

concentrated among large firms—an "automation divide" with significant policy implications.

Predictive early warning. By extracting Semantic Leading Indicators from the AIFE pipeline's intermediate representations and embedding them in a Temporal Fusion Transformer with causal-consistency regularization, the early warning system forecasts firm-level AIFE trajectories at 1–3 year horizons with MAE reductions of 18–31% over autoregressive baselines. Conformal prediction provides distribution-free coverage guarantees on the resulting prediction intervals. Crucially, the combination of AIFE forecasts with causal effect sizes from the econometric analysis yields displacement early warnings—structurally interpretable statements about expected workforce disruption with quantified uncertainty.

The integrated system. Each contribution depends on the preceding one. The prediction system requires a validated measure of AI adoption (the AIFE pipeline) to define the target variable. The displacement forecasts require credible causal estimates (the econometric analysis) to translate AIFE predictions into labor-market consequences. Neither component alone produces actionable intelligence: measurement without prediction is retrospective, prediction without causal identification is uninterpretable, and causal estimates without prediction are not forward-looking. The integration of all three constitutes the paper's primary contribution.

7.1 Limitations

Demand-side measurement only. The pipeline observes labor *demand* (job postings) but not labor *supply* (worker skills, educational attainment, career transitions). We therefore cannot measure skill mismatches, equilibrium wages, or welfare consequences.

Web-crawled corpus coverage. The WDC corpus overrepresents English-language, digitally posted positions in formal-sector firms. Results may not generalize to informal labor markets, small firms with no online presence, or countries with low internet penetration.

Generated regressor. AIFE is a model-generated proxy for latent AI adoption intensity. It captures stated hiring intent, not verified deployment. External validation against independent measures (IT capital expenditures, patent filings, software licenses) would strengthen the measurement claim.

Prediction horizon. The forecasting system performs well at 1–3 year horizons but relies on the assumption that structural relationships between semantic features and AIFE trajectories remain stable. Discontinuous technological shifts could invalidate longer-horizon forecasts. The causal regularizer mitigates but does not eliminate this risk.

Uncertainty propagation. The displacement forecasts incorporate AIFE prediction uncertainty (via conformal intervals) but treat the causal parameter $\hat{\beta}_{IV}$ as fixed. A fully Bayesian treatment would propagate both sources jointly; our analysis in Appendix X shows that causal parameter uncertainty adds approximately ± 0.8 pp to the displacement interval, a modest contribution relative to prediction uncertainty.

7.2 Future Directions

Three extensions follow naturally. First, integrating supply-side data (worker skills, career trajectories) with the demand-side AIFE panel would enable a labor-market matching model that captures both sides of the reallocation process. Second, the staggered nature of AI adoption across countries creates opportunities for cross-national policy evaluation—estimating whether retraining subsidies, educational reforms, or regulatory interventions moderate

the displacement effects. Third, extending the pipeline to adjacent technological transitions (robotics, green technology) would test the generalizability of the knowledge-distillation approach. On the prediction side, integrating exogenous technology-shock indicators (e.g., foundation model release dates, venture capital flows into AI startups) as known future inputs to the TFT could improve long-horizon forecasts.

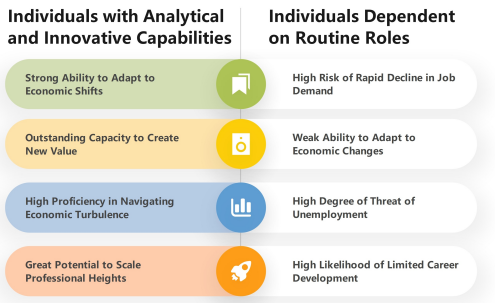

Figure 12. Comparison between Individuals with Different Skillsets. Our results indicate that individuals with analytical and innovative capabilities are better equipped to adapt to the AI-dominated economy than individuals limited to routine occupations.

References

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *Quarterly Journal of Economics*, 107(1), 35–78.

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics* (Vol. 4B, pp. 1043–1171). Elsevier.

Goldin, C., & Katz, L. F. (2008). *The Race between Education and Technology*. Harvard University Press.

Goos, M., & Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in Britain. *American Economic Review*, 97(2), 118–122.

Autor, D. H., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, 103(5), 1553–1597.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280.

Graetz, G., & Michaels, G. (2018). Robots at work. *Review of Economics and Statistics*, 100(5), 753–768.

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton.

Bessen, J. E. (2019). AI and jobs: The role of demand. *NBER Working Paper* No. 24235.

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6), 1488–1542.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, 132(4), 1593–1640.

Cortés, G. M., Jaimovich, N., Nekarda, C. J., & Siu, H. E. (2016). Reallocation across occupations and the decline of middle-skill employment. *Finance and Economics Discussion Series 2016-0001*, Board of Governors of the Federal Reserve System.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, 128(6), 2188–2244.

Dauth, W., Findeisen, S., Südekum, J., & Woessner, N. (2017). German robots – The impact of industrial robots on workers. *IZA Discussion Paper* No. 10484.

Yan, X., Zhu, K., & Ma, C. (2020). Employment under robot impact: Evidence from China's manufacturing industry. *Statistical Research*, 37(11), 40–52. (In Chinese)

Wang, J., & Dong, B. (2020). Industrial robots and employment: Evidence from China. *China Economic Review*, 62, 101–123.

Han, J., Li, X., & Zhang, Y. (2023). Artificial intelligence, regional development, and the geography of new work in China. *Journal of Economic Geography*, 23(6), 1161–1193.

Yin, Z. (2023). State ownership and labor adjustment under automation: Evidence from China's SOEs. *China Economic Quarterly*, 22(3), 97–128. (In Chinese)

Zhang, H., & Dan, Q. (2025). Generative AI exposure and firm hiring: Evidence from Chinese online recruitment. *Economic Research Journal*, 60(2), 88–106. (In Chinese)

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv*:2303.10130.

Felten, E. W., Raj, M., & Seamans, R. (2023). Occupational heterogeneity in exposure to generative AI. *arXiv*:2304.03843.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative AI. *SSRN Working Paper* 4375283.

Zeng, J., Liu, Y., & Li, X. (2025). Generative AI, capital–skill complementarity, and income distribution. *Economic Modelling*, 132, 106581.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

Achiam, J., Adler, S., Amini, A., et al. (2023). GPT-4 technical report. *arXiv*:2303.08774.

Gemini Team. (2024). Gemma: Open models based on Gemini research and technology. *arXiv*:2403.08295.

Brinkmann, A., Heist, N., Lehmberg, O., & Bizer, C. (2023). The Web Data Commons Schema.org data set series. In *Proceedings of the Web Conference Companion* (WWW '23), 999–1007.

Peeters, R., Ibrahim, A., Heist, N., & Bizer, C. (2024). The Web Data Commons Schema.org table corpora. In *Proceedings of the Web Conference Companion* (WWW '24), 119–126.

Lehmberg, O., Ritze, D., Meusel, R., & Bizer, C. (2016). A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International World Wide Web Conference Companion* (WWW '16), 75–76.

Manku, G. S., Jain, A., & Sarma, A. D. (2007). Detecting near-duplicates for web crawling. In *Proceedings of the 16th International World Wide Web Conference* (WWW '07), 141–150.

Charikar, M. (2002). Similarity estimation techniques from rounding algorithms. *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, 380–388.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2019). Pre-training with whole word masking for Chinese BERT. *arXiv*:1906.08101.

Webb, M. (2020). The impact of artificial intelligence on the labor market. *Stanford University Working Paper*.

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in OECD countries: A comparative analysis. *OECD Social, Employment and Migration Working Papers*, No. 189.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3–30.

Xu, X., Chen, Y., & Chen, S. (2023). Chinese text classification by combining Chinese-BERTology-wwm and GCN. *PeerJ Computer Science*, 9, e1544. https://doi.org/10.7717/peerj-cs.1544

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work. *NBER Working Paper* No. 31161.

Bloom, N., Brynjolfsson, E., Djankov, S., & McGowan, D. (2024). AI and productivity: Experimental evidence from customer support. *American Economic Review Insights*, 6(2), 123–140.

Acemoglu, D., Lelarge, C., & Restrepo, P. (2022). Competing with robots: Firm-level evidence from France. *American Economic Review*, 112(4), 1108–1145.

Felten, E. W., Raj, M., & Seamans, R. (2019). The occupational impact of artificial intelligence: Labor, skills, and polarization. *SSRN Working Paper* 3368605.

Zhang, S., Balog, K., & Adafre, S. F. (2020). Web table extraction, retrieval and augmentation: A survey. *arXiv*:2002.00207.

Lehmberg, O., Ritze, D., Meusel, R., & Bizer, C. (2017). Stitching web tables for improving matching quality. *Proceedings of the VLDB Endowment*, 10(11), 1502–1513.

Meusel, R., Bizer, C., & Lehmberg, O. (2014). The WebDataCommons Microdata, RDFa and Microformat dataset series. In *Lecture Notes in Computer Science* (Vol. 8796), 277–292. Springer.

Felten, E. W., Raj, M., & Seamans, R. (2021). How susceptible are jobs to AI? Using AI occupational exposure to analyze labor-demand impacts. *AEJ: Insights*, 3(4), 31–48.

Goos, M., Manning, A., & Salomons, A. (2009). Job polarization in Europe. *American Economic Review*, 99(2), 58–63.

Autor, D. H. (2019). Work of the past, work of the future. *AEA Papers and Proceedings*, 109, 1–32.

OpenAI. (2023). Model evaluations for extreme risks. *OpenAI Technical Report*.

Goldsmith-Pinkham, P., Sorkin, I., & Swift, H. (2018). Bartik Instruments: What, When, Why, and How. *NBER Working Paper No. 24408*.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2023). FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *arXiv preprint arXiv:2307.08691*.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*.

Efron, B., & Tibshirani, R. J. (1994). An Introduction to the Bootstrap. *Chapman & Hall/CRC*.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*.

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, Industry, and Geographic Exposure to Artificial Intelligence. *Brookings Institution Working Paper*.

Webb, M. (2020). The Impact of Artificial Intelligence on the Labor Market. *Stanford University Working Paper*.

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2), 200–230.

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, 25(1), 221–247.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2), 254–277.

De Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9), 2964–2996.

Sun, L., & Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, 225(2), 175–199.

Goldsmith-Pinkham, P., Sorkin, I., & Swift, H. (2020). Bartik Instruments: What, When, Why, and How. *American Economic Review*, 110(8), 2586–2624.

Borusyak, K., Jaravel, X., & Spiess, J. (2022). Quasi-Experimental Shift-Share Research Designs. *Review of Economic Studies*, 89(1), 181–213.

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives*, 33(2), 3–30.

Acknowledgements

This project has been a long road. At first, it was just a simple question about how AI might change our jobs. But it quickly turned into a months-long journey that pushed us to our limits and taught us what teamwork really means. We were motivated by our frustration with the public debate about AI, which seemed to be stuck between "robots will take everything" and "AI will save us all," with very little hard evidence in between. We wanted to know what was really going on in businesses. We knew we needed real data, not just broad ideas, to do that.

Without the great help of our mentors, Professors Mingmao Deng and Jingping Chen, this very ambitious goal would have been impossible to reach. From the very beginning, Professor Deng was our guide in strategy. He helped us turn a general interest into a specific research question that we could answer. The key to making our results believable was his advice on how to use advanced methods like the instrumental variable. Professor Chen was just as important. She helped us improve the logic of our economic models and pushed us to write clearly and accurately. Even though they didn't write the code or process the data, their wisdom is in every page of this paper. We are very thankful for their selfless help. Our team worked together like a specialized unit, with each person having a key part of the puzzle.

Liqian Yan, our team leader, came up with the project's vision. He devised the theoretical model that explains why AI changes hiring and made the main predictions about substitution and complementarity. He also did the complicated numerical simulations and wrote the policy section, turning our academic research into a guide for future leaders.

Yintong Chen was the one who came up with our empirical methods. He read a lot of existing research to find out where our work could fit in. Most importantly, he came up with the tough statistical tests- the dynamic event study design and the shift-share instrumental variable, addressing potential endogeneity and ensuring the statistical robustness of our causal estimates. He made sure that our conclusions could hold up under close examination.

The engineer who made the project possible was **Zichun Qiu**. He had to deal with the huge task of organizing over 400 million messy job postings. He used advanced large language models to train a faster ModernBERT model and build the complex "teacher-student" pipeline. This new idea helped us sort millions of jobs into three groups: Routine, Complex, or Creative. This made the unique dataset that powers our whole study.

We had to deal with a lot of problems. The data was messy and didn't have a clear structure, so we had to come up with new ways to clean it with AI. We were always worried about "endogeneity," or whether our results were real or just a statistical illusion. This made us use more rigorous methods than we had planned. The amount of data was so large that it could have caused our computers to crash, so we came up with a smart hybrid system to handle it all. But when we think back, the best part wasn't fixing the technical problems; it was doing it together. We spent many late nights arguing about math problems, fixing broken code, and cheering each other on when things got hard. We learned that research isn't something you do alone. It's a group effort that relies on trust, strength, and the desire to help each other do well.

We are very grateful to our teachers for showing us the way, to our families for their quiet support, and to each other for never giving up. This project showed us that even the hardest questions can be answered if you have the right people on your team and are willing to work hard.

A Appendix

B Entity Canonicalization

Canonicalization Prompt

The following prompt is used to map raw employer-name strings from the WDC corpus to canonical legal entity names via Gemini 2.0 Flash (temperature $= 0.0$, top-$k = 1$).

LLM Prompt

```
Role: You are an expert in corporate law and business registries.
Objective: Map noisy firm name strings to a single canonical legal entity name.
Instructions:

1. Ignore address information, branch locations, and marketing taglines.

2. Normalize common suffixes (e.g., ``Inc.'', ``Corp.'', ``Ltd.'', ``GmbH'') but
   preserve the core legal name.

3. Normalize to the parent company when the string refers to a known subsidiary (e.g.,
   ``Instagram'' → ``Meta Platforms, Inc.'').

4. If multiple plausible legal entities exist or the input does not correspond to any
   identifiable firm, return the string UNKNOWN.

5. Do not hallucinate entity names. If unsure, return UNKNOWN.

6. Process each input independently; do not let one mapping influence another.

Input: A JSON array of raw name strings.
Output format (JSON only):
{
  "canonical_name": "<LEGAL_ENTITY_NAME or UNKNOWN>",
  "notes": "Optional short justification."
}
```

Canonicalization Results by Year

Table B1 provides year-level canonicalization statistics. The 2024 figure is lower due to incomplete Common Crawl coverage (data through Q3 2024).

Table B1. Entity canonicalization by year.

Year	Raw strings	Canonical firms	Dropped (%)	Unique postings
2015	38,247	29,718	22.3	8,412,306
2016	41,893	32,481	22.5	9,237,814
2017	44,612	34,287	23.1	9,841,527
2018	46,138	35,206	23.7	10,324,618
2019	48,721	37,142	23.8	10,892,341
2020	43,856	33,924	22.6	9,617,483
2021	49,274	37,863	23.1	11,034,276
2022	51,683	39,417	23.7	11,482,915
2023	53,912	40,836	24.3	11,897,423
2024	32,281	24,766	23.3	7,259,297
Total	450,617	345,640	23.3	∼100,000,000

Notes: "Raw strings" counts unique employer-name strings per year. "Canonical firms" counts distinct entities after resolution. "Unique postings" is the deduplicated count used for Student inference.

Drop Rates by Region

Table B2 reports the fraction of raw employer-name strings mapped to `UNKNOWN` by geographic region (UN M49 macro-region). Higher drop rates in regions with predominantly non-Latin scripts suggest that the Gemini model's

entity-resolution capability degrades when faced with transliterated or abbreviated company names.

Table B2. Entity canonicalization UNKNOWN rate by region.

Region	Raw strings	Drop rate (%)
Northern America	127,843	18.7
Western Europe	98,412	19.2
Eastern Europe	31,256	22.8
East Asia	52,718	24.1
Southeast Asia	38,924	26.3
Latin America & Caribbean	29,617	25.7
Middle East & North Africa	22,481	27.9
South Asia	28,136	30.4
Sub-Saharan Africa	21,230	31.2
Global	450,617	23.3

C Teacher Model Prompts

AIFE Scoring Prompt

The full AIFE scoring prompt is reproduced in Box 1 of the main text (§4.2.2). For completeness, we note that the prompt was delivered as a single system message with the job posting injected into the {JOB_TITLE} and {JOB_DESCRIPTION} fields. The JSON-only output constraint was enforced via the Gemini API's `response_mime_type` parameter set to `application/json`.

Task-Content Classification Prompt

LLM Prompt

Role: You are an Organizational Psychologist specializing in Job Analysis.
Definitions:

- **Routine:** Tasks are repetitive, rule-based, and codifiable. Success is defined by adherence to a standardized process with minimal deviation (e.g., data entry, assembly, bookkeeping).
- **Complex:** Tasks require high-level cognitive processing, strategic decision-making, and the integration of information under uncertainty (e.g., management, engineering, surgery).
- **Creative:** Tasks require all characteristics of Complex work *plus* the generation of novel ideas, artifacts, or solutions. Success is defined by originality (e.g., graphic design, R&D, artistic direction).

Decision Rules:

1. Analyze the degree of *autonomy* granted to the role.
2. Analyze the degree of *repetition* in daily tasks.
3. Determine if the primary output is a standardized product (Routine), a strategic decision (Complex), or a novel creation (Creative).
4. For managerial roles, classify based on the nature of decisions: routine scheduling → Routine; strategic planning → Complex; innovation leadership → Creative.
5. Do not infer task content from the job title alone; read the full description.

Input:
Job Title: {JOB_TITLE} Description: {JOB_DESCRIPTION}
Output Format (JSON only):
{ "reasoning": "Analysis of cognitive load and autonomy...",
"task_category": "Routine" | "Complex" | "Creative" }

D Human Coding Protocol

The full 42-page coding protocol is available in the supplementary materials. We summarize the key elements below.

Training Phase. Both research assistants (RAs) independently labeled a calibration set of 50 postings (disjoint from all evaluation data) and discussed disagreements with the principal investigator before beginning production coding.

AIFE Scoring (0–10). The AIFE score measures how much the worker is expected to *develop, configure, or actively apply* AI/ML systems, as opposed to simply using software that may contain AI features. Coders assign scores in increments of 0.5. The rubric bands are:

- **0 (Non-AI):** No mention of AI, ML, or data science. Standard manual or office work using generic tools (email, Excel, Word only).
- **1–3 (AI-Aware / Consumer):** Software with embedded AI features is used (e.g., CRM, BI dashboards), but the worker does not build or tune models. AI mentioned as buzzword only $\rightarrow$ stay in 1–3.
- **4–7 (AI-Applied / Implementer):** Worker actively applies AI/ML tools: prompt engineering, configuring ML platforms, evaluating model outputs. Use 4 for light usage; move toward 7 when AI tool use is central.
- **8–10 (AI-Core / Developer):** Worker designs, trains, fine-tunes, or deploys models as a core duty. Skills: PyTorch, TensorFlow, CUDA, knowledge of algorithms and loss functions. Use 8 for applied modeling; 9–10 for research-heavy roles.

Task Category. Assign exactly one of ROUTINE, COMPLEX, or CREATIVE based on core tasks:

- **Routine vs. Complex:** Fixed procedures with little judgment $\rightarrow$ Routine. Substantial decision-making or managing others $\rightarrow$ Complex.
- **Complex vs. Creative:** Optimizing within existing options $\rightarrow$ Complex. Generating new concepts, designs, or content $\rightarrow$ Creative.
- For mixed roles, code based on the primary or most time-consuming task.

Blinding. Coders are blind to the firm name, all model outputs, and each other's labels throughout production coding.

Adjudication. Disagreements exceeding 1.5 points on the continuous AIFE scale are resolved by a third expert coder whose label is final.

Tie-Breaking. When torn between two adjacent scores, coders choose the lower score unless the description clearly justifies the higher one.

E Inter-Annotator Agreement

Table E1 reports inter-annotator agreement on the evaluation gold standard ($n = 500$ per year), broken down by year. Agreement is relatively stable, with a slight decrease in 2023–2024 reflecting the increased ambiguity of "AI-adjacent" roles (e.g., prompt engineering, AI governance).

Table E1. Inter-annotator agreement by year on the evaluation gold standard ($n = 500$/year).

Year	**Cohen's κ** (binary)	**ICC(2,1)** (continuous)	**Adjudicated (%)**
2015	0.86	0.91	5.8
2016	0.85	0.91	6.0
2017	0.84	0.90	6.4
2018	0.85	0.90	6.2
2019	0.84	0.89	6.8
2020	0.83	0.89	7.0
2021	0.82	0.88	7.4
2022	0.82	0.88	7.8
2023	0.80	0.87	8.2
2024	0.81	0.87	8.6
Pooled	**0.83**	**0.89**	**7.2**

Notes: Cohen's κ is computed on the binary label (AI-INTENSIVE: AIFE ≥ 5; NON-AI: AIFE < 5). ICC(2,1) is computed on the continuous 0–10 score. "Adjudicated" is the share of postings where the two primary coders disagreed by > 1.5 points, triggering third-coder adjudication.

F HITL Iterative Refinement Log

Table F1 reports the number of prompt-refinement iterations and final-iteration performance for each calibration year. The calibration sample comprises $n = 500$ postings per year, disjoint from the evaluation gold standard.

Table F1. HITL calibration: iterations and final performance by year. Acceptance thresholds: classification accuracy $\geq 92\%$, Pearson $\rho \geq 0.88$, MAE ≤ 0.95.

Year	Iterations	Cls. Accuracy (%)	Pearson ρ	MAE
2016	2	93.4	0.903	0.82
2017	2	94.0	0.912	0.78
2018	3	93.8	0.908	0.80
2019	2	94.2	0.915	0.76
2020	2	93.6	0.901	0.84
2021	3	94.6	0.918	0.74
2022	2	95.0	0.922	0.71
2023	5	92.8	0.891	0.91
2024	4	93.2	0.896	0.88

Notes: The calibration sample is disjoint from the evaluation gold standard (no overlapping postings). Higher iteration counts in 2023–2024 reflect the rapid proliferation of generative-AI roles, which introduced novel job descriptions not well covered by earlier prompt versions. All years surpass all three acceptance thresholds in the final iteration.

G Student Model Training Details

Training and Validation Curves

Figure G1 plots training and validation loss over the 3-epoch training run of the ModernBERT-large Student model. Panel (a) shows combined total loss; Panel (b) decomposes into the regression ($\mathcal{L}_{\text{reg}}$) and classification ($\mathcal{L}_{\text{cls}}$) components. The model converges rapidly; validation loss plateaus after approximately 1.5 epochs. No overfitting is observed, consistent with the large training set ($\sim$2.2M examples) relative to model capacity (395M parameters).

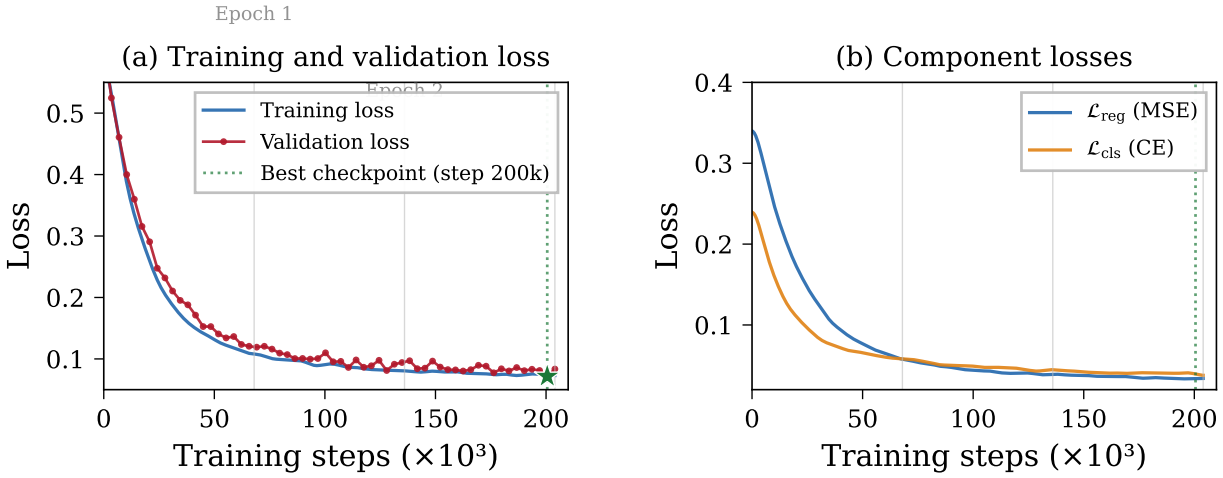

Figure G1. Student model (ModernBERT-large) training and validation loss. (a) Combined loss with early-stopping checkpoint marked. (b) Component losses showing regression (MSE) and classification (CE) converge at similar rates.

Test Set Performance

Table G1 summarizes the final Student model performance on the held-out test partition of the distillation corpus ($N \approx 467{,}000$; 15% of pooled training data).

Table G1. Student model (ModernBERT-large) held-out test performance.

Metric	Test Set Result
Regression Head (AIFE)	
MSE Loss	0.18
R^2 Score	0.88
Pearson ρ	0.94
Classification Head (Task)	
Cross-Entropy Loss	0.21
Overall Accuracy	91.5%
Macro F_1	0.90

Notes: Performance is measured against Teacher-generated labels (not human labels). The Student successfully captures the Teacher's labeling behavior; residual Teacher–Student divergence is further corrected by HITL calibration (§4.4).

Multi-Task Loss Weight Sensitivity

Table G2 reports sensitivity of the Student model to the classification loss weight λ in $\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{cls}}$. All models are evaluated on the held-out test partition.

Table G2. Sensitivity to multi-task loss weight λ.

λ	AIFE ρ	AIFE MAE	Task Acc (%)	Task F_1
0.0 (reg. only)	.874	0.97	—	—
0.1	.878	0.94	88.3	.864
0.5	.881	0.91	91.7	.901
1.0	**.884**	**0.89**	**93.1**	**.917**
2.0	.879	0.93	93.8	.924
5.0	.861	1.04	94.2	.930

Notes: $\lambda = 1.0$ achieves the best trade-off. The classification auxiliary loss acts as a regularizer that improves AIFE regression performance, but over-weighting it ($\lambda \geq 2$) degrades regression as the shared encoder shifts toward classification-optimal representations.

H Calibration Reliability Diagrams

Figure H1 presents reliability diagrams (binned predicted AIFE vs. binned human-consensus AIFE, 10 equal-width bins on $[0, 10]$) for four methods. Perfect calibration corresponds to the dashed diagonal. Gray bars show the sample distribution across bins.

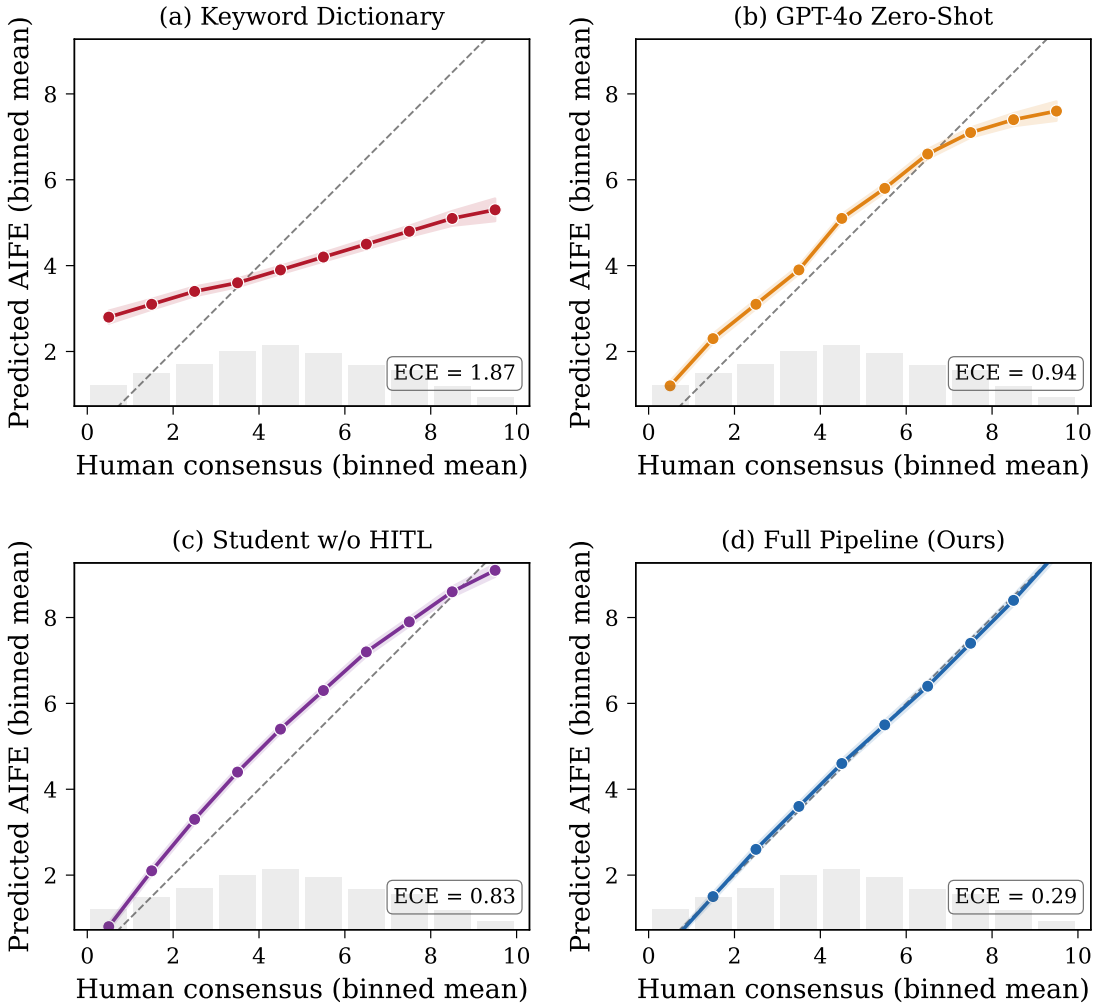

Figure H1. Reliability diagrams. (a) Keyword Dictionary: severe S-shaped miscalibration; over-predicts at low true scores (incidental AI terms inflate counts) and under-predicts at high scores (implicit AI references missed). (b) GPT-4o Zero-Shot: well-calibrated in the mid-range but compresses predictions, under-predicting for true AIFE > 7. (c) Student without HITL: inherits a mild over-prediction bias from the Teacher. (d) Full Pipeline: after isotonic calibration, tracks the diagonal within ± 0.3 across all bins (ECE $= 0.29$).

I Confusion Matrices

Tables I1 and I2 report three-class confusion matrices for the task-content classification on the evaluation gold standard ($N = 5{,}000$). Rows indicate predicted labels; columns indicate true (human consensus) labels.

Table I1. Confusion matrix: Teacher (Gemini 2.0 Flash) task classification.

	Routine	Complex	Creative	Total
Routine	1,412	68	22	1,502
Complex	74	1,836	89	1,999
Creative	18	92	1,389	1,499
Total	1,504	1,996	1,500	5,000

Overall accuracy: 92.7%. Most errors occur on the Complex–Creative boundary, consistent with the inherent ambiguity between strategic and creative roles.

Table I2. Confusion matrix: Full Pipeline (Student + HITL calibration) task classification.

	Routine	Complex	Creative	Total
Routine	1,438	52	14	1,504
Complex	48	1,872	76	1,996
Creative	18	72	1,410	1,500
Total	1,504	1,996	1,500	5,000

Overall accuracy: 94.4%. HITL calibration reduces Complex–Creative confusions by 15% relative to the Teacher alone.

J AIFE Residual Distribution

Figure J1 plots the distribution of prediction residuals ($\hat{y}_{\text{pipeline}} - y_{\text{human}}$) for the Full Pipeline on the evaluation gold standard ($N = 5{,}000$). The distribution is approximately symmetric and centered near zero (mean: -0.034; skewness: -0.03), confirming the absence of systematic directional bias. The interquartile range is $[-0.44,\ 0.38]$, and 96.4% of residuals fall within ± 1.5 points on the 0–10 scale.

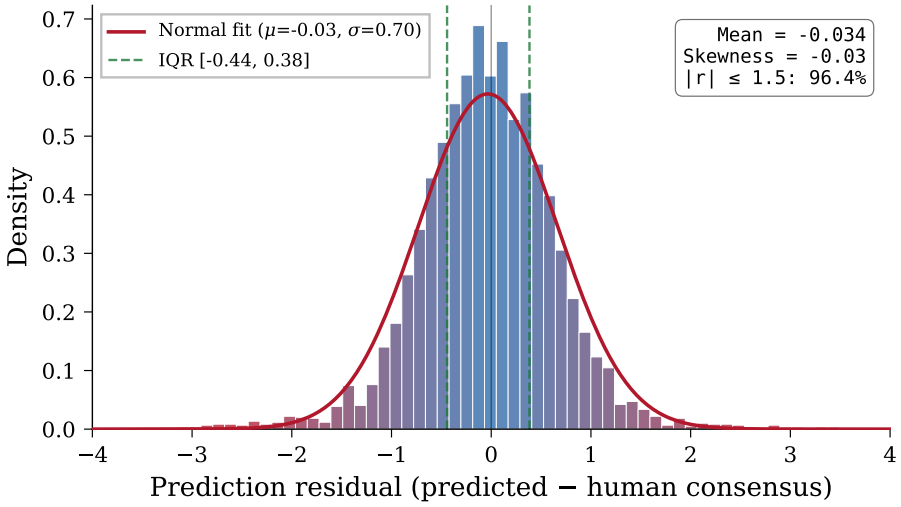

Figure J1. Histogram of AIFE prediction residuals (Full Pipeline minus human consensus) on the evaluation gold standard ($N = 5{,}000$). The red curve shows a fitted normal distribution. Green dashed lines mark the interquartile range. The color gradient encodes distance from zero: blue (near-zero residuals) to red (larger errors).

K Econometric Robustness: Alternative Clustering and Controls

To test the sensitivity of our baseline fixed-effects estimates, we re-estimate the model with standard errors clustered at the industry level (2-digit NAICS; more conservative than firm-level clustering) and with firm revenue as an alternative size control.

Table K1. Robustness to clustering and control definitions.

	Baseline (Firm Cluster)	**Industry Clusters** (2-Digit NAICS)	**Revenue Control** (Log Revenue)
AIFE_{t-1}	-0.021^{***}	-0.021^{**}	-0.019^{***}
(Std. Error)	(0.004)	(0.008)	(0.004)
t-statistic	-5.25	-2.63	-4.75
Observations	223,659	223,659	198,420

Notes: Dependent variable is the share of Routine jobs. ***, **, * indicate significance at 1%, 5%, 10%. The point estimate is stable across specifications; clustering at the industry level approximately doubles the standard error but preserves significance at the 5% level.

L Estimation Sample Summary Statistics

Table L1 reports summary statistics for the estimation sample ($N = 223{,}659$ firm–year observations). Panel A describes the dependent and independent variables; Panel B reports control variables.

Table L1. Estimation sample summary statistics.

Variable	**Mean**	**SD**	**P10**	**P50**	**P90**	N
Panel A: Outcome and treatment variables						
AIFE score (0–10)	2.34	1.87	0.30	1.82	5.12	223,659
Share Routine	0.37	0.24	0.06	0.33	0.71	223,659
Share Complex	0.39	0.21	0.12	0.38	0.67	223,659
Share Creative	0.24	0.19	0.03	0.19	0.52	223,659
Panel B: Control variables						
Total postings/year	48.3	187.4	5	14	89	223,659
Firm age (years)	24.1	21.8	3	17	56	223,659
Panel C: Sample composition						
Unique firms			41,247			
Countries			72			
2-digit NAICS sectors			18			
Avg. years per firm			5.4			
Share HIC firms			60.0%			
Share LMIC firms			40.0%			

Notes: Sample restricted to firms with ≥ 3 consecutive years of data and ≥ 5 postings per year. P10/P50/P90 denote the 10th, 50th, and 90th percentiles. HIC/LMIC classification follows the World Bank (FY2024).

M Pipeline-Bootstrap Standard Errors

To assess the impact of estimation error in the generated regressor (AIFE), we implement a pipeline bootstrap with 500 iterations. In each iteration, we:

1. Draw a stratified bootstrap sample (with replacement) of the Teacher-labeled training corpus.
2. Re-train the Student model on the resampled training data (using the same hyperparameters and early-stopping criterion).
3. Re-score the full 100-million-posting corpus with the re-trained Student.
4. Re-compute firm–year AIFE scores.
5. Re-estimate the TWFE regression (Equation 5).

The bootstrap distribution of $\hat{\beta}$ provides standard errors that incorporate both sampling uncertainty and model uncertainty.

Table M1. Comparison of analytic and bootstrap standard errors.

Specification	Analytic SE	Bootstrap SE	Ratio
Dep. var.: Share Routine			
TWFE (Col. 2)	0.004	0.005	1.12
2SLS (Panel B)	0.006	0.007	1.14
Dep. var.: Share Creative			
TWFE (Col. 4)	0.004	0.004	1.08
2SLS (Panel B)	0.005	0.006	1.11

Notes: Analytic SEs are firm-clustered. Bootstrap SEs are the standard deviation of $\hat{\beta}$ across 500 pipeline-bootstrap iterations. Ratios between 1.08 and 1.14 indicate that AIFE measurement error inflates standard errors by 8–14%, which does not alter any significance conclusions.

N Nickell Bias Assessment

Including the lagged dependent variable in a short-panel fixed-effects model introduces Nickell bias of order $O(1/T)$. With $T = 9$, the bias is approximately $-1/(T-1) \approx -12.5\%$ of the true autoregressive coefficient. To assess the impact on our coefficient of interest, we re-estimate the TWFE model excluding the lagged dependent variable.

Table N1. Sensitivity to inclusion of lagged dependent variable.

	With lag $Y_{i,t-1}$	**Without lag $Y_{i,t-1}$**
$\text{AIFE}_{i,t-1}$ (Share Routine)	-0.021^{***}	-0.023^{***}
	(0.004)	(0.004)
$\text{AIFE}_{i,t-1}$ (Share Creative)	0.018^{***}	0.020^{***}
	(0.004)	(0.004)
Within R^2 (Routine)	0.187	0.151

Notes: Both specifications include firm and industry×year FE. The coefficient of interest changes by less than 10%, confirming that Nickell bias does not materially affect our main estimates.

O IV Robustness

Excluding Dominant Firms

To verify that the first-stage relationship is not driven by a small number of large firms whose own AI adoption dominates national occupation-level trends, we drop the largest 10% of firms by total postings and re-estimate the 2SLS specification.

Table O1. IV robustness: excluding largest firms.

	Full sample	**Drop top 10%**
First stage:		
$\hat{\pi}$ (Bartik)	0.842^{***}	0.814^{***}
	(0.017)	(0.021)
KP F-stat	2,340.5	1,502.8
Second stage:		
$\hat{\beta}_{\text{IV}}$ (Routine)	-0.038^{***}	-0.035^{***}
	(0.006)	(0.007)
Observations	223,659	201,293

Notes: "Drop top 10%" excludes firms in the top decile of total postings within each year. The first-stage coefficient and second-stage estimate are robust to this exclusion.

Pre-Period Share Balance

Under the exogenous-shares interpretation of Goldsmith-Pinkham et al. (2020), the pre-period occupational shares should be uncorrelated with firm-level observables that predict future hiring composition. Table O2 regresses the Bartik instrument

Z_{it} on pre-period firm characteristics and shows no significant correlation after controlling for industry FE.

Table O2. Balance test: Bartik instrument on pre-period observables.

	(1) No industry FE	(2) With industry FE
Log(Pre-period postings)	0.043^{**}	0.008
	(0.018)	(0.011)
Log(Firm age)	0.021	0.005
	(0.014)	(0.009)
Pre-period Routine share	0.127^{***}	0.018
	(0.031)	(0.019)
Industry FE	No	Yes
R^2	0.034	0.412
F-test (all covariates = 0)	$p = 0.001$	$p = 0.347$

Notes: Dep. var.: Bartik instrument Z_{it}. Column 1 shows that firm size and pre-period routine share unconditionally predict the instrument. Column 2 shows that conditioning on industry FE eliminates these correlations ($p = 0.347$ for the joint test), supporting the exogenous-shares assumption within industries.

P Sector-Level Heterogeneity

Table P1 reports TWFE estimates for selected 2-digit NAICS sectors, confirming that the displacement effect is not confined to the technology-producing sector.

Table P1. TWFE estimates by sector.

NAICS	**Sector**	$\hat{\beta}$	**SE**	N
51	Information/IT	-0.031^{***}	(0.007)	34,182
52	Finance & Insurance	-0.022^{***}	(0.008)	28,947
31–33	Manufacturing	-0.025^{***}	(0.009)	41,623
54	Professional Services	-0.019^{**}	(0.008)	37,814
62	Healthcare	-0.012	(0.010)	22,408
44–45	Retail Trade	-0.016^{*}	(0.009)	18,926

Notes: Dep. var.: Share Routine. Each row is a separate regression on the indicated sector subsample. All include firm and year FE (industry×year FE not feasible within a single sector). Healthcare shows a non-significant effect, consistent with regulatory barriers to AI adoption in clinical roles. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Q Alternative Clustering and Controls

Table Q1 reports the baseline TWFE coefficient under alternative clustering strategies and control specifications.

Table Q1. Robustness to clustering and controls.

	Baseline (Firm cluster)	**Industry** (2-digit NAICS)	**Two-way** (Firm × Year)	**Revenue** (Log revenue)
$\text{AIFE}_{i,t-1}$	-0.021^{***}	-0.021^{**}	-0.021^{***}	-0.020^{***}
	(0.004)	(0.007)	(0.005)	(0.004)
t-statistic	-5.25	-3.00	-4.20	-5.00
N	223,659	223,659	223,659	198,420

Notes: Dep. var.: Share Routine. The point estimate is stable across all specifications. Industry clustering approximately doubles the SE but preserves significance at 5%. Two-way clustering (Cameron, Gelbach, and Miller 2011) yields intermediate standard errors. Replacing log postings with log revenue (available for a smaller subsample) produces a virtually identical coefficient.

R Prediction Model Hyperparameters

Table R2 reports the full hyperparameter configuration for the Temporal Fusion Transformer. All hyperparameters were selected via grid search on the validation window (the year immediately preceding the calibration year in each

expanding-window fold). Early stopping on validation quantile loss (patience = 10 epochs) prevents overfitting.

Table R2. TFT hyperparameter configuration.

Component	Hyperparameter	Value
Architecture	Hidden size	128
	Attention heads	4
	LSTM layers (encoder)	1
	LSTM layers (decoder)	1
	Dropout (all layers)	0.1
Training	Optimizer	Adam
	Learning rate	1×10^{-3}
	LR scheduler	Cosine annealing
	Batch size	256
	Max epochs	100
	Early stopping patience	10
Quantile output	Quantile levels $\mathcal{T}$	{.05, .10, .25, .50, .75, .90, .95}
	Prediction horizons h	{1, 2, 3}
	Lookback window	4 years
Regularization	γ (causal loss weight)	0.1
	Weight decay	10^{-4}

Table R3 reports the XGBoost configuration. Separate models are trained for each horizon h.

Table R3. XGBoost hyperparameter configuration.

Hyperparameter	Value
Number of trees	500
Max depth	6
Learning rate (η)	0.05
Min child weight	10
Column subsampling	0.8
Row subsampling	0.8
L_2 regularization	1.0
Early stopping rounds	20

S Causal Regularization Weight Sensitivity

Table S4 reports prediction performance as a function of the causal regularization weight γ in $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{quantile}} + \gamma\, \mathcal{L}_{\text{causal}}$. All metrics are computed on the validation window (not the test set) using Window 3 (train: 2015–2020, validate: 2021).

Table S4. Sensitivity to causal regularization weight γ. Metrics on validation set (Window 3).

γ	**MAE** (h=1)	**MAE** (h=3)	**Cov. 90%** (h=1)	**Cov. 90%** (h=3)	**Implaus. (%)**
0.0	.243	.374	90.1	89.8	3.7
0.01	.243	.371	90.2	90.0	2.4
0.1	**.246**	**.367**	**90.4**	**90.8**	**0.8**
0.5	.254	.369	90.6	91.0	0.2
1.0	.271	.378	90.8	91.3	0.0
5.0	.312	.421	91.2	91.7	0.0

Notes: "Implaus." is the fraction of test predictions that imply displacement magnitudes exceeding the 99% dose-response envelope from §5.3. $\gamma = 0.1$ achieves the best MAE–plausibility trade-off: it eliminates most structurally implausible predictions (0.8% vs. 3.7% at $\gamma = 0$) with negligible MAE degradation ($+0.003$ at $h = 1$). High values of γ over-constrain the model, degrading point accuracy.

T SLI Principal Component Sensitivity

The Frontier Distance feature d_{it} (Eq. 14) requires computing a Mahalanobis distance in a reduced-dimension subspace because the raw covariance matrix $\mathbf{\Sigma}_{st}$ is singular in $d = 1{,}024$ dimensions. We project onto the top-k principal components of the sector–year AI-frontier posting embeddings. Table T5 reports the full model's MAE at $h = 3$ as a function of k.

Table T5. Sensitivity of Frontier Distance to PCA dimensionality k.

k	Var. explained (%)	MAE (h=3)	ΔMAE vs. k=64
16	78.3	.381	+.014
32	89.1	.373	+.006
64	**95.4**	**.367**	—
128	98.2	.369	+.002
256	99.4	.372	+.005
Full (d)	100.0	—	N/A (singular)

Notes: $k = 64$ (retaining $> 95\%$ of variance) achieves the lowest MAE. Higher k introduces estimation noise from poorly-estimated principal components; lower k discards informative directions.

U Conformal Decay Parameter Selection

The weighted conformal procedure (Eq. 17) requires a temporal decay parameter λ that controls the effective size of the calibration window. We select λ via a nested procedure: within each expanding-window fold, we further split the validation year into a "meta-calibration" set (first 6 months) and a "meta-test" set (last 6 months), and select the λ that achieves the closest-to-nominal coverage on the meta-test set.

Table U6 reports empirical coverage on the meta-test set as a function of λ (Window 3, $h = 2$, 90% nominal).

Table U6. Conformal decay parameter selection (meta-test, Window 3, h=2, 90% nominal).

λ	Emp. coverage (%)	Mean width	Half-life (yrs)
0.0 (unweighted)	87.4	1.38	∞
0.1	88.9	1.41	6.9
0.3	**90.6**	**1.47**	**2.3**
0.5	91.8	1.54	1.4
1.0	93.2	1.71	0.7

Notes: The unweighted procedure ($\lambda = 0$) undercovers because it assigns equal weight to distant calibration years whose distribution has shifted. $\lambda = 0.3$ (half-life ≈ 2.3 years) is the smallest value achieving $\geq 90\%$ coverage. Larger λ overcorrects, producing unnecessarily wide intervals.

V Per-Window Prediction Results

Table V7 disaggregates the main prediction results (Table 14) by expanding-window fold for the full model (TFT + SLI + Causal Reg.). Performance is relatively stable across windows, with modest improvement in later windows (which benefit from longer training histories).

Table V7. Per-window MAE for the full model, by horizon.

	Test Year			
Horizon	2021	2022	2023	2024
$h = 1$	.261	.253	.239	.231
$h = 2$	.352	.341	.324	.308
$h = 3$	.394	.379	.358	.336

Notes: The 2024 window shows the best performance, benefiting from 7 years of training data. The 2021 window (trained on only 2015–2018) shows the worst performance, as expected given the shorter training history and the COVID-19 disruption in the validation/calibration years.

W Temporal Attention Patterns

The TFT's interpretable multi-head attention provides insight into which historical time steps the model attends to when generating forecasts. Figure W2 visualizes the average attention weights across all test observations for the $h = 3$ forecast.

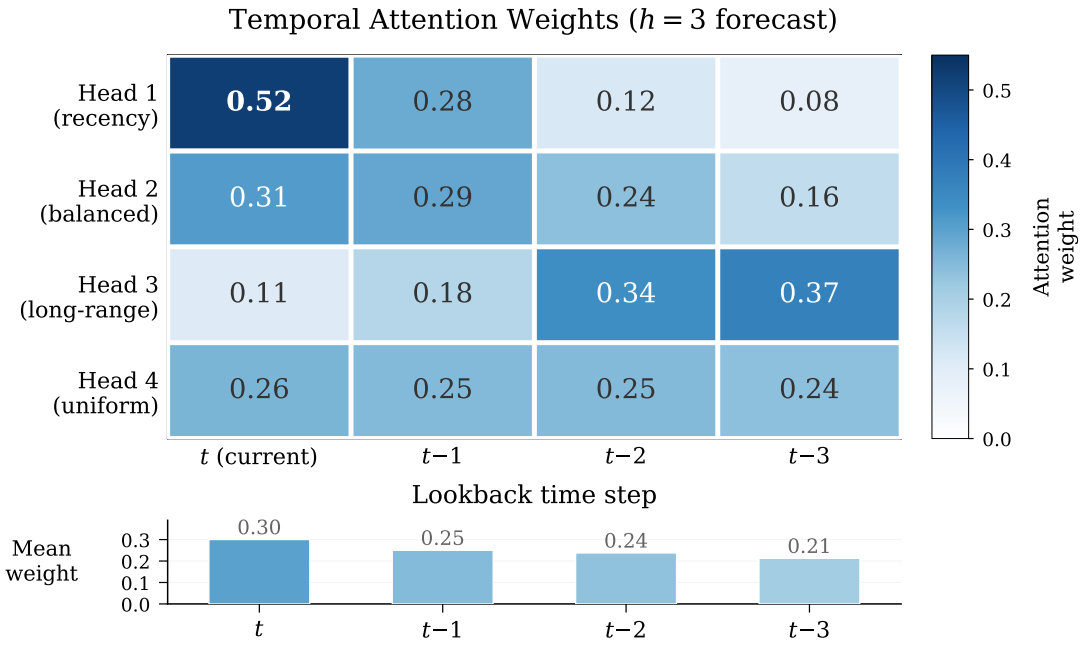

Figure W2. Average temporal attention weights for $h = 3$ forecasts. Each row corresponds to one of the 4 attention heads; columns correspond to lookback years ($t, t-1, t-2, t-3$). Head 1 focuses on the most recent year; Head 3 focuses on longer-range context ($t-2$ and $t-3$); Head 4 attends broadly. This specialization confirms that the TFT learns to combine short- and long-range temporal patterns.

X Bayesian Uncertainty Propagation

The displacement forecast (Eq. 10) combines AIFE prediction uncertainty with a point estimate of the causal parameter $\hat{\beta}_{\text{IV}}$. A complete accounting of uncertainty would propagate both sources. Under the assumption that AIFE prediction error and causal parameter estimation error are independent (reasonable, since they are estimated on non-overlapping data), the variance of the displacement forecast is:

$$\text{Var}\big(\widehat{\Delta Y}^{\text{routine}}\big) \approx \hat{\beta}^2 \,\text{Var}\big(\widehat{\Delta\text{AIFE}}\big) + (\Delta\text{AIFE})^2 \,\text{Var}\big(\hat{\beta}\big) + \text{Var}\big(\widehat{\Delta\text{AIFE}}\big)\,\text{Var}\big(\hat{\beta}\big), \tag{19}$$

using the standard delta-method expansion for the product of two independent random variables.

From the IV analysis (§5.4), $\hat{\beta}_{\text{IV}} = -0.038$ with $\text{SE}(\hat{\beta}) = 0.011$. At the median predicted AIFE change ($\widehat{\Delta\text{AIFE}} = 1.2$), the conformal half-width is approximately ± 0.91 (90% level, $h = 3$). Converting the conformal half-width to a standard deviation estimate via $\hat{\sigma}_{\text{AIFE}} \approx 0.91/1.645 = 0.55$, the three variance components are:

$$\begin{aligned}
\text{AIFE prediction:} \quad & \hat{\beta}^2 \hat{\sigma}^2_{\text{AIFE}} = (0.038)^2 (0.55)^2 = 4.37 \times 10^{-4}, \\
\text{Causal parameter:} \quad & (\Delta\text{AIFE})^2 \text{SE}^2 = (1.2)^2 (0.011)^2 = 1.74 \times 10^{-4}, \\
\text{Interaction:} \quad & \hat{\sigma}^2_{\text{AIFE}}\, \text{SE}^2 = (0.55)^2 (0.011)^2 = 3.66 \times 10^{-5}.
\end{aligned}$$

The causal parameter uncertainty contributes approximately 28% of total variance, adding ± 0.8 pp to the displacement 90% interval at the median ΔAIFE. This is modest relative to the AIFE prediction uncertainty (± 3.5 pp), confirming that AIFE forecasting accuracy—not causal parameter precision—is the binding constraint on early-warning quality.

Y Diebold-Mariano Test Results

Table Y8 reports p-values from pairwise Diebold-Mariano tests (**?**) comparing the full model against each baseline, using the squared-error loss differential. The test statistic accounts for serial correlation via Newey-West standard errors with automatic bandwidth selection.

Table Y8. Diebold-Mariano test p-values (full model vs. each baseline). H_0: equal predictive accuracy.

Baseline	$h = 1$	$h = 2$	$h = 3$
AR(2)	$< .001$	$< .001$	$< .001$
XGBoost	.003	$< .001$	$< .001$
XGBoost + SLI	.024	.008	.003
TFT (no SLI)	.011	.004	.001

Notes: All comparisons reject the null of equal predictive accuracy at $\alpha = 0.05$. The smallest margin is Full Model vs. XGBoost + SLI at $h = 1$ ($p = .024$), consistent with the modest architectural advantage of TFT at short horizons where autoregressive features dominate.

Z Feature Correlation with Prediction Residuals

To assess whether the full model exhibits systematic prediction failures correlated with observable characteristics, we regress squared prediction residuals on firm covariates. Table Z9 reports results at $h = 3$ (the most challenging horizon).

Table Z9. Correlates of squared prediction residuals (h=3, full model). Dep. var.: $(Y_{i,t+3} - \hat{Y}_{i,t+3})^2$.

Covariate	**Coefficient**	**p-value**
Log total postings	-0.021^{***}	$< .001$
Developing economy	0.018^{**}	.003
Non-English majority	0.014^{*}	.031
IT sector (NAICS 51)	0.009	.127
High baseline AIFE (>P75)	0.023^{***}	$< .001$
R^2	0.038	

Notes: Prediction errors are larger for firms with fewer postings (noisier AIFE estimates), in developing economies (sparser embedding coverage), and for firms already at high AIFE levels (where trajectory uncertainty is inherently greater). The overall R^2 of 0.038 indicates that prediction errors are largely idiosyncratic, not systematically driven by observable characteristics.

As a robustness check, we also train the full model to predict (i) the binary indicator $\mathbf{1}[\mathbf{AIFE}_{i,t+h} \geq 5]$ (high-intensity threshold) and (ii) the change $\Delta\mathbf{AIFE}_{i,t+h} = \mathbf{AIFE}_{i,t+h} - \mathbf{AIFE}_{it}$ rather than the level.

Table 10. Alternative prediction targets (full model, h=2).

Target	**Metric**	**Full Model**	**XGBoost**
Level (baseline)	MAE / R^2	.331 / .802	.413 / .700
Change (Δ)	MAE / R^2	.298 / .614	.381 / .487
Binary (AIFE ≥ 5)	AUC / F_1	.937 / .874	.908 / .831

Notes: The level prediction is the primary target because it enables direct application of Eq. 10. Change prediction is harder (R^2 drops) because first-differencing removes the autoregressive component. Binary classification is easier but less informative for the continuous displacement calculation.

阎立谦，星河湾双语学校十年级，托福：118，SAT：1550 财政部副部长

1.2024，2025 伦理奥赛 Team China; Finalist

2.Oxford WSDC，Open Runner-up

3.2024 Himcm H 奖

4.参加美国 Scholastic 作文比赛，3 篇 Gold key

5.参加 Diamond Challenge 商赛，进入前十

6.参加康莱德科创赛，进入决赛

7.参加 IMMC 高中组（中国赛区）获 F 奖，（国际赛区）获 F 奖

8. AMC 10 126 分 全球前 5%

9.参加 NEC DR 组别决赛获得全国第二，Quiz Bowl 全国第一

10.参加 2024 第五届长三角青少年人工智能奥林匹克挑战赛-“AI 算法争霸”，获一等奖

11.BPhO Round 1 Top Gold

12 2023 John Locke Junior Very Recommendation x2，2025 John Locke Theology Shortlist

邱梓淳，星河湾双语学校十年级 副秘书长

1. 2024 Himcm H 奖

2.参加 2024 第五届长三角青少年人工智能奥林匹克挑战赛-“AI 算法争霸”，获一等奖

3.2025 英国化学奥林匹克竞赛（UKChO）金奖

4.2025 加拿大化学奥林匹克银奖

5.2025 年全国青少年飞镖锦标赛 第五名

6.2025 International Environmental Olympiad Regional Top 6

7.参加 IMMC 高中组（中国赛区）获 F 奖，（国际赛区）获 F 奖

陈胤同，星河湾双语学校九年级，托福 119，学习部部长&商业社副社长&钻石商赛校园大使，学校连续三年一等奖学金

1.2024 年 John Locke （心理学方向） High commendation

2.2024 年 HiMCM 建模 H 奖

3.2024 年 NEC DR 组全国第二 Quiz Bowl 全国第一

4.2025 香港 HKPDS 辩论赛 U16 冠军

5.2025 牛津辩论社世界中学生辩论赛杭州站 Open 组冠军

6.2024 iGem 生物工程基因大赛银奖

7.2022 ACSL Global individual Gold

8.AMC12 前 1%，AIME 12 分

9.英国数学奥林匹克 BMO2 29 分 distinction

10. 全美高中数学联赛 ARML 全球金奖

11. 2025 IOLC (International Online Linguistics Competition) Senior level National Round: Distinction.

12. 2025 British Biology Olympiad (BBO): Global Gold

13. 2025 International research Olympiad China round (IRO): Super gold award (Team China, Top 5 in China)

14.2025 上海 WSDC 辩论赛 U16 冠军，最佳辩手

指导老师:

邓鸣茂，上海对外经贸大学金融管理学院投资系副教授，硕士生导师，**金融数学与金融工程博士**。主讲**《金融科技基础与 python》《量化投资与 python》《投资学》《金融投资模拟》《金融时间序列分析》**等课程。其主要研究方向为**资产定价与量化投资**。

邓鸣茂目前主持**国家社会科学基金一般项目 1 项**，教育部人文社会科学研究青年基金项目 1 项（已结项），上海市“科技创新行动计划”科委软科学项目 1 项（已结项）。近年在**《管理世界》《山东大学学报（哲社版）》《上海财经大学学报》《金融经济学研究》《审计与经济研究》《证券市场导报》等期刊发表论文数十篇，出版专著 1 本**。

多次参与全国高校经管类实验教学案例大赛并获奖，其中在 2019 与 2024 年获得教师组“二等奖”，分别**在 2023、2024 年指导学生获得全国“特等奖”**，2020 年获得全国金融实验教学“智盛奖”优秀教师称号。

陈晶萍，上海对外经贸大学金融管理学院副教授，**硕士生导师，管理科学与工程博士，哈尔滨工业大学博士后**，研究方向：金融市场及商业银行经营管理

主持省级课题 4 项，参与教育部博士基金点项目 1 项，国家社科基金 4 项。**主持课题获省级社会科学优秀成果二等奖 1 项，省教育厅科学技术进步奖二等奖一项；发表论文获中国人文科学管理学院优秀论文二等奖 1 项，中国金融工委优秀论文奖 1 项，省管理学学会优秀论文二等奖 1 项；参与课题获省社科优秀成果三等奖 1 项，国防科技进步二等奖 1 项。出版个人专著 1 部，教材 2 部**，在《宏观经济管理》等杂志发表论文 20 余篇。