Identifying AI Jobs in Online Job Postings:

Existing Versus Newly Developed LLM-Based Approaches
and A Dataset on Demand for AI Jobs among Russell 1000 Firms, 2018–2025

Hanwen Shi
hwshi@umd.edu

Anil K. Gupta
agupta@umd.edu

Waverly Ding
wding@umd.edu

Kunpeng Zhang
kpzhang@umd.edu

Robert H. Smith School of Business
The University of Maryland
College Park, MD 20742

Manuscript-in-Progress
November 2025

Identifying AI Jobs in Online Job Postings:

Existing Versus Newly Developed LLM-Based Approaches
and A Dataset on Demand for AI Jobs among Russell 1000 Firms, 2018–2025

Abstract

Research Summary. Artificial intelligence (AI) is spreading rapidly across industries, yet firms differ widely in adoption intensity, making measurement of AI investment a core empirical challenge. Investment in AI-skilled human capital is the most widely used proxy. This paper compares existing Lightcast (Burning Glass) skills-taxonomy based methods with large language model (LLM) approaches and introduces a fine-tuned LLM that markedly improves classification performance, raising precision by about 30% and F1 score by about 18% without sacrificing recall or computing efficiency. Using this method, we construct and publicly release a dataset of quarterly AI-skilled labor demand for Russell 1000 firms from 2018 to 2025. Preliminary analyses show that AI-related jobs exhibit a clear upward trend and substantial variation across sectors, regions, and occupations. We also release our code and models and offer methodological guidance for applying LLMs in business research, emphasizing the value of task-specific fine-tuning.

Managerial Summary. As firms race to embed artificial intelligence (AI) into their operations, managers need reliable ways to gauge their own and competitors' AI investments. Hiring for AI-related skills serves as a practical proxy, yet commonly used Lightcast (Burning Glass) skills-taxonomy based methods often misclassify postings, limiting strategic insight. We show that a fine-tuned large language model (LLM) provides far more accurate and efficient classification, sharply reducing false positives and improving key performance metrics. Using this method, we construct and publicly release a dataset of quarterly AI-skilled labor demand for Russell 1000 firms from 2018 to 2025, revealing a clear upward trend and substantial variation across industry sectors, geographic regions, and occupations. These insights can potentially help managers benchmark AI talent strategies, assess emerging competitive pressures, and make more informed capability-building decisions.

Keywords: AI, artificial intelligence, human capital, job posting, large language models

1. INTRODUCTION

Artificial intelligence (AI) is emerging as one of the most transformative technologies in history (West & Allen, 2018). Empirical evidence increasingly shows that it is a general-purpose technology, an extremely rare class that has included the steam engine, electricity, and computing (Eloundou, Manning, Mishkin, & Rock, 2024; Goldfarb, Taska, & Teodoridis, 2023). What distinguishes AI from prior technologies, however, is the remarkable speed of its diffusion. An OpenAI report notes that, as of July 2025, 700 million users were interacting with ChatGPT on a weekly basis; as the report states, "For a new technology, this speed of global diffusion has no precedent" (Chatterji et al., 2025).

As has been true of other general-purpose technologies, despite AI's broad applicability, there exists considerable variation in the pace at which different sections of the economy are adopting it (Chatterji et al., 2025; Shao, Zope, Jiang, Pei, Nguyen, Brynjolfsson, & Yang, 2025). Not surprisingly, economists and management scholars have been increasingly interested in studying how and why sectors, firms, and occupations vary in their adoption of AI and, with or without such variations, the consequences of AI adoption (e.g., Acemoglu, 2025; Babina, Fedyk, He, & Hodson, 2024; Brynjolfsson, Mitchell, & Rock, 2018; D'Acunto, Prabhala, & Rossi, 2019; Felten, Raj, & Seamans, 2021; Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2022).

Existing studies on AI adoption can be classified into three broad streams. The first focuses on individual users as the unit of analysis, examining the impact of AI use on the quality and efficiency of the user's work. For example, Noy and Zhang (2023) investigate ChatGPT's effect on writing by college educated professionals; Otis, Clarke, Delecourt, Holtz, and Koning (2024) assess the impact of AI-assisted advice on the revenues and profits of Kenyan entrepreneurs; and Brynjolfsson, Li, and Raymond (2025) study the impact of introducing a generative-AI based conversational

assistant to customer service agents. The second stream includes studies that focus on occupations (such as locomotive engineers, actuaries, accountants, mathematicians, et al.) as the unit of analysis and measure heterogeneity in the AI exposure of different occupations (e.g., Brynjolfsson et al., 2018; Felten et al., 2021). Building on the O*NET database (https://www.onetonline.org), these studies use various methods to estimate the AI exposure of "Tasks" and "Detailed Work Activities" associated with each occupation and then aggregate these estimates upward to the occupational level. Extending this logic, Acemoglu (2025) aggregates across occupations to estimate the macro-impact of AI on the overall economy.

The third stream, which relates to the objective of this paper, takes the firm as the unit of analysis, focusing on how firms differ in their use of AI and/or the consequences of such heterogeneity (Acemoglu, Autor, Hazell, & Restrepo, 2022; Alekseeva, Azar, Giné, Samila, & Taska, 2021; Babina et al., 2024; Goldfarb, Taska, & Teodoridis, 2020, 2023; Makridis & Alterovitz, 2024). A methodological challenge for this line of research is the measurement of AI adoption by a large population of firms over an extended period. Because of the embedded nature of AI, it is challenging to measure AI adoption by studying the firm's output of products and services. For example, Apple's iPhone camera uses AI to analyze multiple frames before the shutter is pressed and to combine them into a single, high-quality image (Mah, 2021); this embedded use of AI would be nearly impossible for external observers to decode. This is true as well for AI embedded in other products and services, be they in banking, insurance, retail, or any other sector. It is also nearly impossible for scholars to decode whether an organization is using AI in its internal processes. For competitiveness reasons, firms tend not to reveal much if any information about internal use. As a result, beyond case studies and anecdotal accounts (e.g., Meister & Chandrasekhar, 2021), internal use if any also remains a black box for external observers.

Faced with these challenges in decoding the use of AI to enhance products, services, or processes, researchers have tended to rely on the inputs to AI use. Two of the most important such inputs are: computing power (typically proxied by data center capacity) and AI-skilled human capital. Measuring investments in or leasing of data center capacity also suffers from the limitations of private and impacted information. While a handful of companies (such as Amazon, Microsoft, Google, and Meta) do reveal the details of their investments in data centers, similar information is unavailable for most other firms, including giants such as Walmart and JP Morgan Chase.

Consequently, investment in AI-skilled human capital has emerged as the dominant indicator of firm-level AI technology adoption. Studies typically operationalize this investment using online job postings for AI-skilled roles (Acemoglu et al., 2022; Alekseeva et al., 2021; Babina et al., 2024; Goldfarb et al., 2020, 2023; Makridis & Alterovitz, 2024). An alternate, albeit less common, approach is to estimate the stock of AI-skilled employees in a database of LinkedIn or similar profiles (Babina et al., 2024; Goldfarb, He, & Teodoridis, 2025; Hampole, Papanikolaou, Schmidt, & Seegmiller, 2025). As Babina et al. (2024) note, the two measures, when scaled appropriately, are highly correlated.

In this paper, we focus on the identification of AI-skilled jobs in online job postings, the dominant approach employed in the literature examining firms' adoption of AI. In particular, we *(1) evaluate the performance of existing approaches used to decode whether or not a job posting requires AI skills, (2) propose a new method for identifying AI-related jobs, based on fine-tuning one of the world's most powerful large language models (LLMs), and (3) apply the new LLM-based method to the universe of all online job postings by Russell 1000 firms since 2018-Q1 through 2025-Q3. We also share code and models and make publicly available, for each firm per quarter, the count of both AI-skilled job postings and the full set of job postings, thereby creating a foundation*

for future research by other scholars.

2. IDENTIFYING JOBS REQUIRING AI SKILLS IN ONLINE JOB POSTINGS

2.1 The Informational Value of Online Job Postings

Firms routinely post job openings on their career webpages to attract qualified applicants, rendering job postings a rich and publicly observable trace of firm behavior. As a data source, online job postings offer several important advantages. First, researchers have access to comprehensive coverage through major data aggregators such as Lightcast[1] (formerly Burning Glass; referred to as LBG henceforth) and LinkUp, both of which have been scraping the career pages of employers since circa 2007 and currently add data on about 20 million new job postings each year. Second, the data are highly granular. Each job posting includes detailed metadata, such as posting and removal dates, employer name and ticker symbol if any, geographic location of the job, NAICS code to identify the employer's industry and sector, and the occupational O*NET code. Third, unlike corporate websites, press releases, and even earnings calls, job postings tend not to suffer from much window dressing. Employers post job openings on their career webpages to attract suitable job candidates. Embellishments that might attract a greater number of less suitable candidates would be self-defeating.

Given their granularity, comprehensiveness, and informational value, online job postings have attracted growing attention from social scientists. For example, Azar, Marinescu, Steinbaum, and Taska (2020) use job postings data to examine the degree of occupation-based labor market concentration in each commuting zone. Hershbein and Kahn (2018) use data on online job postings to study if recessions accelerate the shift of production towards routine-biased technologies and

[1] The data aggregator formerly known as Burning Glass Technologies has undergone two name changes – first to Emsi Burning Glass, and more recently to Lightcast. For the sake of easier links to earlier literature, we use the term Lightcast-Burning Glass (or LBG) in this paper.

find that they do. Deming and Kahn (2018) study variations in skill demand for professionals and, focusing particularly on cognitive and social skills, report positive correlations between each skill and external measures of pay and firm performance. Gutiérrez, Lourie, Nekrasov, and Shevlin (2020) report that job postings are informative to investors; more specifically, investors react positively to growth in the number of job postings, especially when firms are likely to be hiring for growth rather than replacement. Castilla and Rho (2023) study the gendering of job postings in the online recruitment process and report that "in practice, employers' efforts to simply tweak the language of recruitment messages do not matter much for gender equality and diversity" (p. 6912). Huang, Pacelli, Shi and Zou (2024) report that disclosing information about corporate culture in job postings is helpful to employers in attracting and retaining workers. Lee and Kim (2024) analyze job postings by startups to study the phenomenon of "scaling" and find that early scalers face greater odds of failure. More recently, Ham, Hann, Rabier and Wang (2025) analyze online job postings by accounting firms and find that a stated demand for auditors' cognitive and social skills correlates positively with audit quality.

2.2 Existing Approaches to Identifying AI Jobs in Online Job Postings

A central challenge for scholars interested in estimating the number of job postings for AI-skilled roles is determining whether a given posting truly requires AI-related skills.

Keyword-based identification, whether applied to job titles or to the full text of job postings, is highly unreliable. Consider two actual postings from Microsoft. The first is for a "Logic Design Engineer." While the job title is silent about AI-relatedness, the raw text of the job posting states that "We are looking for a Logic Design Engineer to work in the dynamic Microsoft Artificial Intelligence System on Chip (AI SoC) Silicon team" and the required or preferred qualifications include "Experience working on Artificial Intelligence (AI)/Machine Learning (ML) SoCs." It is

clear that this job requires AI skills. The second example pertains to a job posting for a "Modern Work & AI Digital Sales Specialist." Note that the job title includes the keyword "AI." However, the raw text of the job posting states that "As a Modern Work & AI Digital Sales Specialist, you will be a solution sales expert within our enterprise sales organization, working closely with our most valued customers" and the required or preferred qualifications include "4+ years of technology-related sales or account management experience." It is clear that this is a sales role that does not require AI skills. Relying on AI-related keywords in the job title would have resulted in a false negative for the first job and a false positive for the second. While relying on keywords in the full text of the job posting would have yielded a true positive for the first job, it would have yielded a false positive for the second.

Because of such limitations, nearly all studies to date have relied on a Skills Taxonomy provided by Lightcast-Burning Glass (LBG-ST, for short). Table 1 contains a summary of these studies.

Insert Table 1 about here

Applying its algorithmic models to the text of each job posting, LBG tags each posting with a set of skills[2]. LBG-ST currently includes over 33,000 distinct skills (see https://lightcast.io/open-skills), among which AI-related skills form a subset. Since LBG's algorithms for skill tagging remain private, scholars have no recourse but to assume that these algorithms are reliable.[3] Given the lack of any alternatives, this assumption seems valid, especially since the LBG algorithms are

[2] LinkUp, the other aggregator of job postings data, caters primarily to quant hedge funds and does not create any skills taxonomy. Instead, it makes the complete raw text of all job postings available to the licensee. It is up to the licensee to code these data.

[3] On its website (https://lightcast.io/open-skills), Lightcast notes that the skills taxonomy is developed through crowdsourcing from economists and labor market analysts. Alekseeva et al. (2021) provide additional visibility into the Lightcast-Burning Glass method: "BGT [Burming Glass Technologies] identifies AI-related skills based on the presence of words and phrases commonly associated with the knowledge of AI (e.g., artificial intelligence, machine vision, deep learning, image processing, speech recognition, etc.) and AI-related software/packages (e.g., IBM Watson, TensorFlow, Pybrain, Random Forests, ND4J, etc.) in skill requirements part of a job description" (p. 3).

likely to yield more accurate results than those from applying keywords directly to either job titles or the complete raw text of job postings.

Except for Babina et al. (2024), studies simply treat a job as an AI job if LBG-ST associates it with at least one AI-related skill. Even then, there are subtle variations in the approaches taken by different scholars.

- Goldfarb et al. (2020) treat a job as an AI job if it is labeled by LBG-ST as belonging to one of the following skill clusters: "Artificial Intelligence," "Machine Learning," and "Natural Language Processing."
- Alekseeva et al. (2021) treat a job as an AI job if it is labeled by LBG-ST as requiring any of 71 specific AI skills.
- Squicciarini and Nachtigall (2021) treat a job as an AI job if it contains at least two AI-related keywords in the LBG-ST list of skills, along with the proviso that one of the two need not be software-related.
- Acemoglu et al. (2022) treat a job as an AI job if it is labeled by LBG-ST as requiring any of 33 specific AI skills.
- Goldfarb et al. (2023) treat a job as an AI job if it is labeled by LBG-ST as requiring at least one skill in the broad skill cluster of "Machine Learning."
- Makridis and Alterovitz (2024) treat a job as an AI job if it is labeled by LBG-ST as requiring any of 172 specific AI skills.

These variations in classification approaches – even when the authors rely directly on AI skill labeling by LBG-ST – leave open the possibility that the findings of any study might be sensitive to alternative classification approaches.

Babina et al.'s (2024) classification approach is more complex than the straightforward reliance

on LBG-ST in the above-noted studies. They adopt a three-step approach: (1) Using the entire database in LBG-ST, compute the AI-relatedness of each skill as its co-occurrence frequency with the following core AI skills: "Artificial Intelligence (AI)," "Machine Learning (ML)," "Natural Language Processing (NLP)," and "Computer Vision (CV)." Across all skills, this AI-relatedness can fall anywhere in the 0-1 range. (2) For each job posting, compute its mean AI-relatedness across all skills associated with that job in LBG-ST. This computation results in a continuous AI-relatedness measure across all job postings. (3) Transform the continuous AI-relatedness measure into a binary indicator by classifying a job as an AI job if the mean exceeds the threshold of 0.1.

Despite its sophistication, Babina et al.'s (2024) approach also raises questions. As they note, based on the co-occurrence measure in Step 1, the skill "TensorFlow" has an AI-relatedness of 0.9. However, in reality, its AI-relatedness should be 1.0, since Google Brain invented TensorFlow as a software library explicitly for the training and interference of neural networks (Oremus, 2015). Other measures also raise questions. As they note in Supplemental Materials (Table A1), the skill "Supervised Learning (Machine Learning)" has an AI-relatedness of only 0.695 and the skill "Neural Networks" has an AI-relatedness of only 0.640, respectively. In reality, both skills have a definitive AI-relatedness of 1.0. Moreover, as per their Supplemental Materials (Table A3), their method yields an overall AI-relatedness score of 0.315 for a job titled "Machine Learning Engineer" – well below face-valid expectations of 1.0. Since 0.315 is above the 0.1 cutoff, Babina et al. (2024) do classify it as an AI job. However, applied to several hundred million job postings, the above-noted issues suggest that the accuracy of their classification method remains unclear. Babina et al.'s (2024) method also raises the question of how sensitive their results might be to cut-off levels above 0.1. On the face of it, it appears that 0.1 might be a low bar to treat any job with an AI-relatedness above this threshold as an AI job. It is not inconceivable that this method may lead to many false

positives in the set of jobs classified as AI jobs. In this method, classification errors can also creep in due to their assignment of equal weight to every skill – whether "Neural Networks" or "Leadership" - when computing the mean AI-relatedness across all skills associated with that job in LBG-ST.

In sum, traditional approaches for identifying AI jobs in online job postings suffer from the following key limitations: First, the algorithms used by LBG to create their Skills Taxonomy remain private and undisclosed. As a result, the sophistication of these algorithms remains unknown. Based on changes over time in the total number of skills in LBG-ST, it also appears that LBG keep refining their algorithms on an ongoing basis. The lack of transparency about these changes makes it difficult to know how sensitive the classification of AI jobs may be to such algorithmic changes. Second, even among the authors who rely most directly on AI skill labeling by LBG-ST, there are subtle differences in how each set of authors chooses to use these labels. It remains unknown how sensitive the classification of AI jobs may be to each study's unique choices. Third, Babina et al.'s (2024) method yields AI-relatedness measures of many skills (e.g., "TensorFlow," "Supervised Learning (Machine Learning)," "Neural Networks") and jobs (e.g., "Machine Learning Engineer") that are at odds with ground-level reality. Their method also assigns equal weight to every skill when computing the mean AI-relatedness of a job and relies on an AI-relatedness cutoff level of a seemingly low 0.1 to classify a job as an AI-related job. These methodological features could potentially introduce significant classification errors.

2.3 Opportunity for New LLM-Based Approaches

Google's invention of the "Transformer" architecture in 2017 is viewed by scientists in the machine learning (ML) field as a turning point in the evolution of ML technologies (Vaswani et al., 2017; Wolf et al., 2020). Building upon this breakthrough, OpenAI's Generative Pre-trained Transformer

(GPT) series of Large Language Models (LLMs) powered the creation and release of ChatGPT in November 2022. Similar LLMs have since emerged from Google (Gemini series), Anthropic (Claude series), Meta (Llama series), Mistral (Mixtral and Mistral series), Alibaba (Qwen series), DeepSeek (DeepSeek series), and others.

LLMs continue to become more powerful. Today's LLMs are far more powerful than those of two years ago and thus likely to be more powerful than LBG's algorithms used to create their Skills Taxonomy. This development presents an opportunity for scholars interested in identifying AI jobs in online job postings to apply LLM-based approaches directly to the raw text of all job postings instead of needing to rely on the Skills Taxonomy created by the undisclosed LBG algorithms.

As we demonstrate in this paper, the application of LLMs to classify job postings requires scholars to make several design choices:

- Whether to rely on closed-source LLMs (e.g., ChatGPT, Claude, Gemini) or open-source ones (e.g., Llama, Qwen, DeepSeek, Mistral)? Closed-source models are often more powerful but also tend to be slower and more expensive in terms of usage costs.
- What power LLM to use? In any given LLM series, users may choose between lower- vs. higher-parameter variants (e.g., Qwen 3-4B vs. Qwen 3-14B). Larger-parameter models typically deliver superior performance but at the expense of increased compute-time and higher usage costs.
- Which prompting strategy to use for a base LLM, "zero-shot prompting" or "few-shot prompting"? Few-shot prompting can increase model performance with a handful of examples. This can be useful in enabling researchers to rely on faster and less expensive smaller models.
- Whether or not to fine-tune the LLM? Fine-tuning is possible only with open-source models. Fine-tuning an LLM means training it further on a specialized dataset (e.g., human-labeled job

postings). Because of its domain specialization, fine-tuning can potentially close the performance gap between smaller and larger open- or closed-source alternatives.

By carefully combining these design elements, scholars can optimize the trade-offs among performance, computational efficiency, and usage cost, paving the way for scalable and transparent measurement of AI demand in labor markets.

2.4 ML-Based Methods in Strategy Research

In recent years, management and strategy scholars have increasingly incorporated machine learning methods into their research, reflecting what Von Krogh, Roberson, and Gruber (2023) describe as the "enlarging of the methodological toolbox" to analyze vast and varied data (p. 367). Since the public release of LLMs is relatively recent, most prior studies have relied on ML approaches that predate the transformer architecture - the technical foundation of LLMs. In this subsection, we review a selection of such studies to highlight how ML methods have been applied to strategic contexts, with the goal of being illustrative rather than comprehensive.

In one of the early studies using ML methods in strategy research, Choudhury, Wang, Carlson, and Khanna (2019) utilize a synthesis of ML methods to discover the communication styles of CEOs – incorporating both verbal and nonverbal aspects of communication. They also report that communication styles were significantly correlated with M&A outcomes. In another study, Choudhury, Allen, and Endres (2021) implement ML methods to study employee turnover in a large technology company and show that ML models can predict employee turnover far more accurately than logistic regressions can. Choi, Menon, and Tabakovic (2021) apply topic modeling (an NLP technique) to unstructured text data in annual reports to create a new multidimensional measure of firm diversification. They go on to show that this method yields novel insights about firm diversification as compared with the established approach of using quantitative financial data.

Miric, Jia, and Huang (2023) demonstrate how to employ ML tools to classify unstructured text documents and create quantitative variables that capture embedded features and show that ML methods are superior to keyword-based approaches. They then "apply these methods to identify AI-based technologies from all patents in the United States, based on patent abstract text" (p. 492). Luo, Jia, Ouyang, and Fang (2024) use advanced deep machine learning and data fusion methods to uncover the intra- and inter-modal interdependencies embedded in unstructured text, image, and audio data within the same source, such as a video. They use live stream commerce on TikTok as the empirical context and show that multimodal data fusion provides more accurate and time-varying measures of trustworthiness than unimodal data. Harrison et al. (2023) demonstrate the value of using supervised machine learning (SML) to scale human-coded data. To illustrate the value of SML, they "focus on the board leadership context, and specifically on the control and collaboration orientations of the board chair position" (p. 1783). Using the data, they examine relationships among board leadership, firm performance, and CEO dismissals. Rathje, Katila, and Reineke (2024) note that, in the first stage of two-stage matching models, researchers often make somewhat ad hoc, and potentially incomplete, choices regarding likely confounding variables. Using propensity score matching as an example, they showcase the "use of ML in the first stage to determine variables on which to match" (p. 1945) and note that "ML methods [can be] useful for choosing matching variables…[and]…help address many of these limitations" (p. 1947).

Collectively, these studies demonstrate how pre-transformer ML methods, ranging from topic modeling to supervised machine learning, enable management scholars to derive novel insights from unstructured data. Building on this foundation, recent research has begun to explore the methodological potential of generative AI and large language models (LLMs) for strategic and organizational analysis.

Doshi, Bell, Mirzayev, and Vanneste (2025) conduct and compare the results of two studies. In Study 1, they use GPT-4 from OpenAI to generate 60 startup business models (across 10 industries, 2 prompts, and 3 probabilities of succeeding). Using pairwise comparisons, they then evaluate these business models using three different types of evaluators - generative AI, human experts, and human nonexperts. In Study 2, the researchers obtain 60 business models submitted to a business model competition offering a total of USD one million in prize money. Pairwise evaluations were conducted using the same procedure as in Study 1. The researchers find that individual evaluations by LLMs tend to be inconsistent and biased. However, aggregating many pairwise evaluations results in greater agreement between the evaluations by LLMs and those by human experts, and more than that between human experts and human non-experts.

Paralleling the Doshi et al. (2025) study, scholars in fields adjacent to strategy (such as accounting and finance) have also started using LLMs as methodological tools. De Kok (2025) highlights the applications of Generative LLMs for accounting research. He demonstrates that LLMs can detect nonanswers in earnings conference calls, a traditionally challenging task to automate. He also notes that the LLM-based method reduces the nonanswer error rate by 70% relative to existing methods. As another illustrative example, Kim , Muhn, Nikolaev, and Zhang (2025) use LLMs to decode topics within annual reports and note that "based on our testing, [this method] is more flexible, robust and comprehensive relative to the previous generation of topic classification approaches (e.g., LDA topic modeling)" (p. 12).

Building on this line of work, our study examines how LLMs can be leveraged for large-scale classification tasks in strategic research contexts. Specifically, we showcase how LLMs yield far more accurate distinctions between AI and non-AI jobs in online job postings, compared with existing approaches.

2.5 Focus of This Study

Building directly upon the above arguments, this paper aims to accomplish the following:

1. We develop a large human-labeled "ground truth" dataset of job postings, annotated by multiple graduate students with computer science backgrounds and currently enrolled in an analytical STEM master's program.
2. We evaluate different LLMs (and their adaptations such as zero-shot prompting, few-shot prompting, and fine-tuning) against the ground-truth dataset. We present the corresponding results of these tests.
3. We present our conclusions regarding which LLM (and which adaptation of it) emerges as the best in terms of performance, computing efficiency, and usage cost.
4. We compare the performance of existing methods based on LBG-ST and the top-performing LLM-based method for identifying AI jobs in online job postings.
5. We apply the top-performing LLM to the entire universe of all job postings by Russell 1000 firms over the period 2018/Q1-2025/Q3 and make the resulting dataset publicly available to facilitate future research. We also make the code and the models publicly available.

Throughout, we emphasize methodological transparency to promote replicability and to enable future scholars to refine or extend our approach as more powerful LLMs emerge.

3. SELECTION OF METHODS FOR EVALUATION

In this section, we introduce the methods used to identify AI-related jobs and the criteria guiding model selection. Prior empirical studies on jobs requiring AI-related skills have employed a variety of classification strategies. Building on the review in Section 2, we begin by summarizing the most representative methods in the extant literature and then describe emerging base LLM approaches - which, to our knowledge, have not yet been applied in this context - and the fine-tuned LLM method

developed in this study.

3.1 LBG Skill Label Method

The LBG Skill Label method is the most widely used method for identifying AI-related jobs and examining their impact on the labor market (e.g., Acemoglu et al., 2022; Alekseeva et al., 2021; Goldfarb et al., 2020, 2023; Makridis & Alterovitz, 2024; Squicciarini & Nachtigall, 2021). This method relies on a predefined list of AI-related skills labelled by LBG. A job posting is classified as AI-related if it requires at least one of these skills. The primary advantage of this method lies in simplicity: it can be implemented efficiently across large datasets without specialized software or significant computational resources. However, it also carries notable risks of the type discussed in Section 2.2. Given the different ways in which various studies have implemented the LBG Skill Label method (see Table 1), we pick the specific approach implemented in one illustrative study - Acemoglu et al. (2022) - for evaluations in our analysis.

3.2 LBG Skill-Relatedness Method

To refine identification, Babina et al. (2024) developed an AI skill-relatedness measure using the same LBG skill label data but applied differently. As discussed in Section 2.2, they adopt a three-step approach: Using the entire database in LBG-ST, start by computing the AI-relatedness of each skill as its co-occurrence frequency with the four core AI skills: “Artificial Intelligence (AI),” “Machine Learning (ML),” “Natural Language Processing (NLP),” and “Computer Vision (CV).” Then, for each job posting, compute its mean AI-relatedness across all skills associated with that job. Finally, transform the continuous AI-relatedness measure into a binary indicator by classifying a job as an AI job if the mean exceeds the threshold of 0.1. This method also suffers from several key limitations – as we detailed earlier in Section 2.2. Given its relative sophistication, we include the LBG Skill-Relatedness method as implemented by Babina et al. (2024) as well for evaluations

in our analysis.

3.3 Base LLMs

In Section 2.3, we discussed the potential of leveraging LLMs to identify AI-related jobs. Unlike traditional machine learning methods requiring extensive feature engineering, base LLMs (also known as pretrained models) allow researchers to perform classification tasks through natural language prompts. A job posting can be classified by instructing the model (e.g., "Determine whether this job requires AI-related skills") without additional programming or labeled data. This flexibility makes LLMs particularly attractive for large-scale text classification. However, two key challenges arise: (1) selecting appropriate models, and (2) determining optimal interaction methods.

Model selection is particularly complex given the vast number of available LLMs that vary in size, architecture (e.g., from hundreds of millions to hundreds of billions of parameters), and accessibility – from small open-source models (e.g., Qwen 3-0.6B) that can be downloaded to large closed-source models (e.g., GPT-4o) that are accessible only via cloud interfaces. Since evaluating all available models is infeasible, it is critical to identify a representative subset with potential for high performance. To this end, we adopted the following approach for LLM selections.

For closed-source models, we selected OpenAI's *GPT-4o* and Anthropic's *Claude-3.*7, two of the most powerful and publicly accessible models as of June 2025. As for open-source models, we applied three criteria to guide the selection of a meaningful evaluation set. First, we ensure inclusion of leading models from the world's three frontier regions for AI development - the United States, China, and Europe. Second, we focus on several of the most recent publicly available versions (as of June 2025) from each model family. Third, we require the feasibility of inference on a single GPU[4] (48 GB memory or less, e.g., NVIDIA A100 or RTX 4090), enabling fast replicability by

[4] GPUs (Graphics Processing Units) are a type of semiconductor chip that are especially useful for artificial intelligence and other parallel computing tasks.

academic researchers. Applying these criteria yielded 12 generative models from three families: U.S. (Meta): *Llama 3.2-3B-Instruct*, *Llama 3.2-1B-Instruct,* and *Llama 3.1-8B-Instruct*; China (Alibaba): *Qwen 3-14B*, *Qwen 3-8B*, *Qwen 3-4B*, *Qwen 3-1.7B*, *Qwen 3-0.6B, Qwen 2.5-14B-Instruct*, and *Qwen 2.5-7B-Instruct;* and Europe (Mistral): *Mistral v0.3-7B-Instruct, and Mistral Nemo-Instruct-2407*.

The second challenge is determining the optimal way to interact with LLMs. Broadly, two approaches are commonly used: zero-shot and few-shot prompting. In zero-shot prompting, the model is given a direct instruction to perform a task without any examples. In contrast, few-shot prompting, also referred to as in-context learning (Dong et al., 2024), provides the model with a small set of labeled examples - both positive and negative - alongside task instructions. These examples help guide the model's understanding and improve performance, particularly on tasks where the instructions alone may be ambiguous. In our experiments, we evaluate both prompting strategies across the selected models. Full prompt templates for each setting are provided in Appendix 1.

3.4 Fine-Tuned LLMs

Supervised fine-tuning is an effective approach for adapting LLMs to domain-specific tasks, since base LLMs are pretrained primarily on broad, general-purpose corpora and thus optimized for open-domain conversation rather than specialized applications. As a result, without further adaptation, their performance on domain-specific tasks, such as identifying AI-related jobs, can be suboptimal.

The two commonly used paradigms for supervised fine-tuning are: (1) Instruction tuning (Wei et al., 2021): the model is further trained on examples that teach it how to follow task-specific instructions. Each training instance contains a general instruction (e.g., *"Classify whether this job is AI-related"*) paired with the job description (i.e., input), while the output is a natural language

response such as *"Yes, this is an AI-related role."* This process resembles the model's original pretraining but is conducted on more task-specific, instruction-formatted data. A key advantage of this approach is that it retains the model's conversational and generative capabilities while aligning it more closely with the desired task (e.g., classification objective). (2) Discriminative fine-tuning: it modifies the architecture of generative models slightly to perform direct binary classification. Each training instance is also a pair of input-output. The input is the job posting text (without any instruction), and the output is a binary label ("1" =AI-related, "0" = not AI-related). This method is computationally efficient and avoids post-processing of free-form text outputs. In contrast, instruction-tuned LLMs may produce varied or verbose responses (e.g., "Yes, it is likely AI-related because …"), even when explicitly prompted to output only "1" or "0," which increases the effort needed to standardize the results.

Insert Figure 1 about here

In this study, we fine-tune selected base LLMs using both approaches and compare their performance in identifying AI-related jobs. Figure 1 illustrates the overall evaluation process for various LLMs.

4. DATA

In this section, we describe the construction of the evaluation datasets, including the procedures for matching postings across sources, human annotation, and the partitioning of data into training and testing sets.

4.1 Raw Text of Job Postings and LBG Skills Taxonomy

Evaluating the various methods noted above requires a dataset that: (1) contains the full text of every job posting – to develop and evaluate LLM-based approaches and (2) provides, for the same

set of job postings, the skill labels from the LBG Skills Taxonomy – to evaluate existing LBG-ST based approaches. To meet both requirements, we merged job postings data from LinkUp (full text of job postings) and LBG (Skills Taxonomy labels). At the time we began this study, our licensed LBG data contained skill labels but not raw text, necessitating the use of LinkUp for the raw text of job postings.

While scientific research into machine learning dates back to the 1950s, the broader deployment of ML technologies accelerated around 2010–2012, following breakthroughs in deep learning i.e., the use of multi-layered neural networks enabled by large datasets and powerful GPUs. Other major milestones include the open-source release of the software libraries TensorFlow by Google in 2015 and PyTorch by Facebook (now Meta) in 2016, which significantly expanded the practical adoption of neural networks among software engineers, as well as the introduction of the transformer architecture by Google in 2017 (Vaswani et al., 2017). To tap into data on job postings that would account for these AI technologies, we picked a random one-month period (January 2019) subsequent to these developments. Using with January 1, 2019 as the starting date, we matched the two data sources (LBG and LinkUp) using an exact-match procedure based on (1) firm name, (2) U.S. state, (3) job title, and (4) a three-day posting window. We stopped this process upon reaching 200,000 matched job postings, which provides a reasonably large sample for method evaluation.

4.2 Human-Labeled Ground Truth Dataset

Labeling the raw text of all 200,000 job postings was infeasible. We therefore adopted a resource-efficient three-step protocol designed to preserve accuracy.

Step 1 (Dictionary pre-screen). In consultation with ML researchers, we constructed a comprehensive 113-string dictionary (see Appendix 2) and applied it to the full set of 200,000 job postings, partitioning them into two subsets: Seemingly-Not-AI postings (n = 190,230) containing

none of these dictionary keywords, and Seemingly-AI postings (n = 9,770) containing at least one such keyword. The assumption was that, subject to further assessments, the Seemingly-Not-AI subset would be unlikely to contain any AI jobs, whereas the Seemingly-AI subset would likely contain true positives along with many false positives.

Step 2 (Negative-set validation). We randomly selected 2,000 job postings from the Seemingly-Not-AI subset (n = 190,230) and assigned each to two trained human annotators (graduate students with computer science backgrounds, currently enrolled in a STEM master's program). Each annotator was instructed to carefully read the entire text of each job posting and label whether the job requires technical AI skills (see Appendix 3 for annotation instructions). Both annotators identified zero AI jobs in this sample of 2,000 postings. Given this robust result, we deemed it appropriate to treat all 190,230 Seemingly-Not-AI job postings as Truly-Not-AI (i.e., true negatives).

Step 3 (Positive-set adjudication). We assigned each of the 9,770 Seemingly-AI job postings to two trained human annotators. We employed 10 trained annotators (all graduate students with computer science backgrounds, currently enrolled in a STEM master's program), such that each annotator independently labeled about 2,000 job postings. Across the 9,770 job postings, the inter-annotator agreement (IAA) was 0.85 i.e., the annotators concurred on 85% of job postings. The remaining 15% were sent to third annotators, whose labels served as the tiebreaker. This process resulted in the labeling of 2,327 job postings as Truly-AI (i.e., true positives) and 7,443 as Truly-Not-AI (i.e., false positives). This finding confirms that even a comprehensive AI-string dictionary yields substantial false positives.

Based on the above three-step process, we conclude that, on a ground truth basis, of the 200,000 job postings, 2,327 are Truly-AI and 197,673 (= 190,230 + 7,443) are Truly-Not-AI. This human-

labeled dataset serves as the primary benchmark for evaluating both new LLM-based and existing approaches for the identification of AI jobs in online job postings.

4.3 Partitioning into Training and Testing Sets

For approaches that do not require training or fine-tuning, we report performance on the full ground-truth dataset of 200,000 job postings.

For approaches that require fine-tuning, we adopt the conventional 80-20 train-test split widely used in machine learning: 160,000 job postings for fine-tuning and 40,000 for evaluation. To improve learning efficiency and reduce computational costs, we followed a two-step procedure within training: first, using the comprehensive 113-string dictionary discussed earlier, we filtered out the Truly-Not-AI jobs to alleviate severe class imbalance; then, we fine-tuned the models on the remaining Seemingly-AI jobs, which comprise a mix of both true positives and false positives. Training on all 160,000 job postings would have substantially increased computation costs without yielding any performance gains; given the pronounced class imbalance, it could even degrade classifier accuracy. It is also worth noting that, unlike the extensive datasets required for pre-training (e.g., the development of models such as ChatGPT and Claude), fine-tuning tends to be particularly effective in smaller-data regimes (Zhang, Liu, Cherry, & Firat, 2024).

5. EVALUATION RESULTS

In this section, we first outline the metrics used to evaluate performance across the methods discussed in Section 3. We then report comparisons of performance across these methods and conclude with a cost analysis.

5.1 Evaluation Metrics

Each method is evaluated on both predictive performance (measured by precision, recall, accuracy, and F1 score), and efficiency (measured by time cost and monetary cost).

Precision captures the reliability of positive predictions by measuring the proportion of jobs classified as AI-related that are indeed true AI jobs.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

Recall measures the proportion of actual AI jobs that are successfully identified.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Accuracy measures the overall correctness of the classification, reflecting the proportion of both AI jobs and non-AI jobs that are correctly classified.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives} \quad (3)$$

F1 is the harmonic means of precision and recall, providing a balanced metric that accounts for both correctness and completeness. This is particularly useful in AI job classification, where the dataset is imbalanced and precision–recall tradeoffs are critical.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

Given that the ground-truth dataset is highly unbalanced (2,327 AI vs. 197,673 non–AI jobs), overall accuracy is not very informative - a naïve model labeling all jobs as non-AI jobs would still achieve 98.8% accuracy. We therefore focus on precision (the share of true AI jobs among those predicted as AI), recall (the share of true AI jobs correctly identified), and the F1 score, which are more informative measures of classification performance.

Efficiency metrics include the average processing time in seconds required to classify a single job, and the inference cost in USD. Together, these metrics reflect not only predictive performance but also the computational and economic feasibility of different approaches for large-scale applications.

5.2 Performance Comparison Across Base LLMs (Zero-Shot Classification)

Following the two-step procedure described above, we first applied a comprehensive dictionary of AI-related keywords (Appendix 2) to partition job postings into those with and without AI-related terms. Postings without any AI-related keywords were reliably classified as *Truly-Not-AI*, while those containing at least one AI-related keyword may include both *Truly-AI* and *Truly-Not-AI* jobs. The base LLMs were then applied only to the latter subset of job postings with at least one AI-related keyword. This two-step design maintains classification performance while reducing computational and financial costs by approximately 95%, since only about 5% of job postings contain AI-related keywords.

Insert Table 2 about here

Table 2 summarizes performance across models from different families, generations, and parameter scales against the full ground truth dataset of 200,000 job postings. Using the two-step procedure, we conduct five independent runs per model and report mean performance and standard deviation across the five runs. See Appendix 4 for a detailed discussion of consistency-related considerations for LLM models.

There are three main takeaways from this table. First, Qwen 3-14B outperforms both closed-source models (GPT-4o and Claude 3.7) and achieves the highest F1 score (0.791) among all models.

Second, larger models generally tend to outperform smaller ones, though with diminishing marginal returns. For instance, within the Qwen family, performance improves as model size increases (e.g., the 14B model surpasses both 8B and 4B variants in F1 and precision scores). Smaller models often produce unbalanced results: some exhibit high recall but low precision (e.g., Qwen 3-4B), while others achieve better precision than recall (e.g., Llama 3.2-3B-Instruct). The

gains from increasing model size diminish at the upper end. For instance, the improvement from Qwen 3-8B to Qwen 3-14B is relatively modest. Similar patterns emerge in the Llama and Mistral families, where the 7B- or 8B-parameter models achieve results comparable to much larger closed-source models (OpenAI GPT-4o and Anthropic Claude-3.7) and the best-performing zero-shot model (Qwen 3-14B). These findings suggest that, for zero-shot classification of job postings, models with around 7B parameters are able to provide stable and reliable performance.

Third, we observe generational improvements. Qwen 3 outperforms its predecessor Qwen 2.5. The Mistral family also shows generational improvements, though by smaller margins. The Llama family does not yet offer comparable models across multiple generations, as successive releases differ in model size, making direct generational comparisons infeasible.

5.3 Impact of Few-Shot Classification with LLMs

Prior research suggests that few-shot prompting can improve LLM performance (Dong et al., 2022). However, our results show mixed effects when applied to AI-related job classification. Table 3 reports the performance results with few-shot prompting and changes relative to zero-shot prompting.

Insert Table 3 about here

For most models, few-shot prompting generally increases precision as well as consistency (lower standard deviation) but reduces recall. This pattern likely arises because the examples used in the few-shot prompts include typical AI and non-AI jobs, leading the models to adopt a narrower and more conservative decision boundary. As a result, models achieve higher precision (i.e., a greater share of predicted AI jobs are true positives) but lower recall (i.e., under-identifying postings that describe AI work in less conventional terms). Given the heterogeneity in how AI-related roles

are expressed across industries and occupations, a limited set of examples fails to capture the heterogeneity, and this may cause the model to overlook atypical postings. Consequently, few-shot prompting overall does not significantly improve F1 performance and often sacrifices recall while increasing input length and processing time. Therefore, zero-shot prompting is the more efficient prompt strategy for large-scale AI-related job classification.

5.4 Impact of Fine-Tuning Open-Source LLMs

We next fine-tune the two best-performing open-source models (Qwen 3-14B and Qwen 3-8B) and the most cost-efficient model (Qwen 3-0.6B) using two methods: instruction tuning (yielding Qwen 3-14B-IT, Qwen 3-8B-IT, and Qwen 3-0.6B-IT) and discriminative tuning (yielding Qwen 3-14B-DIS, Qwen 3-8B-DIS, and Qwen 3-0.6B-DIS).

Recall that the overall ground-truth dataset consists of 200,000 job postings. Following the standard 80-20 train-test split widely used in machine learning, we used 80% of the dataset for fine-tuning and 20% for the evaluations reported in this section. Here as well, we use the same two-step approach described in Section 5.2 (and applied for the tests in Tables 2 and 3). Additionally, as discussed earlier and depicted in Figure 1, instruction-tuned models, which preserve conversational ability, are evaluated under both zero-shot and few-shot settings. In contrast, discriminatively tuned models are evaluated directly on the binary classification task.

Insert Table 4 about here

Table 4 reports the results. Both fine-tuning approaches substantially narrow the performance gap associated with model size. For instance, as we saw earlier in Table 2, the base Qwen 3-0.6B model initially lags behind Qwen 3-14B and Qwen 3-8B by around 0.70 in F1 score. After fine-tuning, however, both Qwen 3-0.6B-IT and Qwen 3-0.6B-DIS achieve performance comparable to

their larger counterparts. Notably, the performance improvement with fine-tuning is most pronounced for the smaller models. For instance, in the case of zero-shot prompting after instruction tuning, on the F1 score, Qwen 3-0.6B improves by 0.730; in comparison, Qwen 3-8B improves by much smaller margin of 0.075, and the largest model Qwen 3-14B by an even smaller 0.042. This pattern suggests that fine-tuning provides substantial performance gains for smaller models while yielding diminishing returns as model capacity increases.

Across the two fine-tuning approaches, discriminative models perform slightly better than their instruction-tuned counterparts. Few-shot prompting does not yield additional improvements, likely because (i) the few-shot examples offer limited coverage across the heterogeneous job postings spanning various sectors and occupations, and (ii) the fine-tuning instructions were designed for zero-shot inference, providing only general guidance rather than example-based adaptation.

5.5 Cost Analysis

Time and monetary costs depend on hardware configurations and API pricing, which vary over time. To ensure comparability, all open-source models were run on NVIDIA A100 (40GB) GPU via Google Colab ($0.59/hour as of June 2025), while proprietary models were accessed through their respective APIs. Table 5 reports runtime and cost for classifying 10,000 job postings.

Insert Table 5 about here

Among closed-source models, Claude 3.7 is an order of magnitude more expensive than GPT-4o for both zero-shot and few-shot prompting; the data for zero-short prompting are: Claude 3.7 (24,548 seconds, $2.17) versus GPT-4o (4,340 seconds, $1.60). The large discrepancy likely arises because Claude 3.7 is designed for explicit reasoning, generating intermediate outputs that increase usage costs. Within open-source model families (without fine-tuning), cost scales with model size.

Small models (e.g., Qwen 3-0.6B and Llama 3.2-1B-Instruct) complete classification tasks within 1,000 seconds and cost less than $0.15. Runtime and cost for larger models (e.g., Qwen 3-14B and Llama 3.1-8B) roughly double though they remain far below closed-source models.

Few-shot prompting further inflates cost nonlinearly with prompt length and model size - nearly doubling runtime and cost for models above 7B parameters. In contrast, smaller models such as Qwen 3-1.7B, Qwen 3-0.6B, and Llama 3.2-1B exhibit only marginal cost increases, remaining cost-efficient for prompt-intensive tasks.

Fine-tuned models show increases in inference time relative to the base models (about 70% on average) due to parameter-efficient fine-tuning overhead. Between the two fine-tuning strategies, discriminative models are generally faster. This is notably the case for Qwen 3-0.6B-DIS which emerges as the most efficient, with only a 2% increase in inference time over its base model with zero-shot prompting.

From an application perspective, efficiency is critical for processing large-scale job posting datasets. Although the F1 score of Qwen 3-0.6B-DIS is only marginally smaller than that of the best-performing fine-tuned model Qwen 3-8B-IT with zero-shot prompting (0.827 versus 0.831, see Table 4), it delivers over 60% reduction in both inference time and computational cost. *Therefore, Qwen 3-0.6B-DIS represents the optimal balance between performance and cost-effectiveness.*

5.6 Benchmarking the Best LLM Against Existing Methods

Having identified Qwen 3-0.6B-DIS as the optimal model (close to the highest performance across all LLMs, coupled with the lowest costs in terms of both compute time and monetary costs), we benchmark it against the two existing methods widely adopted in prior literature: the LBG Skill Label method (Acemoglu et al., 2022) and the LBG Skill-Relatedness method (Babina et al., 2024).

All three methods are evaluated on the held-out test set of 40,000 job postings that were not used during fine-tuning for Qwen 3-0.6B-DIS.

First, consistent with the procedure used in the tests reported in earlier sections, we apply AI-related keyword pre-filtering to exclude job postings lacking AI-related terms and then use the remaining subset to conduct the classification using each method.

Insert Tables 6.1 about here

Table 6.1 reports the results. The LBG Skill Label method used by Acemoglu et al. (2022) slightly outperforms the LBG Skill-Relatedness method used by Babina et al. (2024) on F1 score (0.684 vs. 0.657) and precision (0.590 vs. 0.535). However, Qwen 3-0.6B-DIS surpasses both by substantial margins. Benchmarked against LBG Skill Label method, Qwen 3-0.6B-DIS improves F1 score by 0.143 and precision by 0.262, with only a modest 0.010 decrease in recall. Existing methods' higher recall comes at the expense of false positives: more than 40% of job postings they label as AI-related are in fact not. By contrast, Qwen 3-0.6B-DIS achieves a more balanced and accurate classification.

Insert Tables 6.2 about here

To maintain comparability with prior literature, we also evaluate all three methods without keyword pre-filtering to maintain consistency with the practice used in the existing studies (see Table 6.2). Using this approach, the best performing LLM, Qwen 3-0.6B-DIS, again outperforms existing methods by significant margins. Relative to the LBG Skill Label method, it improves the F1 score by 0.181 and precision by 0.291, while reducing recall by only 0.010. And, relative to the LBG Skill-Relatedness method, it improves the F1 score by 0.204 and precision by 0.332, while

reducing recall by only 0.049.

It is worth noting that the higher recall achieved by existing methods comes at the expense of false positives. Regardless of whether one uses AI keyword pre-filtering or not, the precision scores of these models reveal that between 40-55% of job postings labeled as AI-related are in fact not. By contrast, Qwen 3-0.6B-DIS achieves a more balanced and superior classification with much fewer false positives.

Insert Figure 2 about here

Figure 2 summarizes performance across both settings, confirming that the performance gains of our proposed method are substantial and robust to different evaluation conditions. Importantly, Qwen 3-0.6B-DIS performs strongly on its own and becomes even more efficient when combined with a dictionary-based prefiltering step.

In Section 2.2, we speculated that the LBG Skill-Relatedness method's cutoff of 0.1 might be a low bar, leading to the classification of many non-AI jobs as AI jobs, and thus prone to a high rate of false positives. To test whether higher thresholds may possibly improve the performance of this method, we conducted threshold sensitivity analyses.

The results of these sensitivity analyses (see Appendix 5) show that, based on F1 scores, the performance of the LBG Skill-Relatedness method degrades rapidly – rather than improves - at higher thresholds, irrespective of whether one uses the two-step approach (with keywords filtering) or the one-step approach (without any keywords filtering). While precision improves, recall drops sharply. We conjecture that this rapid drop in recall (i.e., failing to classify "true AI jobs" as "predicted AI jobs") stems from two major weaknesses of this method discussed earlier in Section 2.2: (1) the lower-than-valid AI-relatedness score assigned to many AI-specific skills (e.g.,

"Supervised Learning (Machine Learning)" and "Neural Networks" have AI-relatedness scores of only 0.695 and 0.640 respectively, even though both have a definitive AI-relatedness of 1.0); (2) the equal weighting of all skills – whether "Neural Networks" or "Leadership" - when computing the mean AI-relatedness across the skills associated with that job in LBG-ST.

6. OPEN-SOURCE DATASET ON RUSSELL 1000 FIRMS (2018/Q1 - 2025/Q3)

To promote transparency, replication, and future research on firm-level AI adoption, we construct and publicly release a large-scale open-source dataset on AI-related hiring among Russell 1000 firms, spanning 2018/Q1 through 2025/Q3. In doing so, we follow the open-data practices exemplified by Miric et al. (2023) and Albert, Eklund, and Tang (2025). Building on the evidence presented in Section 5, we recommend a two-step process to classify AI-related job postings: (1) keyword filtering – first, apply a comprehensive dictionary-based filter to exclude job postings that are clearly unrelated to AI, thereby reducing noise and computational cost. (2) LLM-based classification – next, apply a validated LLM-based classifier such as Qwen 3-0.6B-DIS, which combines high precision and recall with strong overall F1 performance, to the filtered subset.

We apply this two-step method to the complete LinkUp job posting database, which provides near-universal coverage of online job postings across private, public, government, and nonprofit organizations in the United States. To facilitate future research, we will release an open-source version of the dataset containing the firm-level quarterly aggregates for Russell 1000 companies. The dataset will span the period from 2018/Q1 to 2025/Q3 and include the following four key variables: (1) firm identifier – ticker symbol, (2) AI-related job count – number of job postings classified as AI-related in each quarter, (3) total online job posting count – total number of online postings in each quarter, and (4) AI Jobs Intensity - the proportion of AI-related postings to total postings.

The release will also include the AI-related keyword dictionary, the fine-tuned model weights, prompt templates, and comprehensive documentation detailing the data sources, preprocessing steps, and model evaluation procedures. Together, these materials will enable full reproducibility and facilitate comparative or longitudinal studies of AI diffusion, workforce transformation, and firm-level digital strategy.

7. AI JOBS DEMAND IN THE U.S. ECONOMY

In this section, we provide the results of applying the best-performing LLM-based classification method (as discussed above) to determine the extent of AI-related job postings in the U.S. economy –essentially a measure of labor demand for technical AI skills. By design, the goal of this section is descriptive rather than evaluative or theoretical.

Figure 3 depicts the trend in AI-related job postings from 2018/Q1 through 2025/Q3. Figures 4.1 and 4.2 illustrate the distribution of AI-related job postings across NAICS sectors. Figures 5.1 and 5.2 show the geographic distribution of AI-related job postings across the 50 U.S. states and Washington, D.C. Figures 6.1 and 6.2 display the distribution of AI-related job postings across O*NET occupations.[5]

Insert Figures 3 through 6.2 here

The data in these figures yield the following conclusions.

As Figure 3 indicates, the AI Jobs Intensity of the U.S. economy has grown significantly from 0.22% in 2018/Q1 to 0.78% in 2025/Q3. In terms of raw numbers (data not indicated on the figure), there were 8,360 AI-related job postings in 2018/Q1. In contrast, this number was 37,014 in 2025/Q3 - a function of both an increase in AI Jobs Intensity and growth in the total number of job

[5] The authors remain open to sharing other slices of the data.

postings.

The data in Figure 3 are also suggestive of a strong ChatGPT effect. As reported by the media (Davis, 2021) and supported by our data, the Covid pandemic led to a sharp growth in technology hiring during the second half of 2020 through the end of 2021. Over-hiring during this period led to a sharp slowdown during 2022. The data in Figure 3 suggest that OpenAI's release of ChatGPT on November 30, 2022 played a significant role in the rapid growth in demand for AI jobs since mid-2023.

As Figure 4.1 indicates, aggregated over the entire time period 2018/Q1-2025/Q3, five sectors account for bulk (87.05%) of AI job postings in the U.S. economy: Information (23.21%), Professional, Scientific, and Technical Services (21.88%), Finance and Insurance (16.35%), Manufacturing (15.38%), and Retail Trade (9.83%). As Figure 4.2 indicates, the first four sectors also exhibit AI Jobs Intensity levels significantly above the average of 0.39% for the entire U.S. economy. Retail Trade has a below-average AI Jobs Intensity (0.24%); however, the relative high share of AI jobs for this sector is a function of the fact that this sector accounts for about 1 in 6 of all job postings in the U.S. economy, second only to the overall numbers for Health Care and Social Assistance.

As Figure 5.1 indicates, aggregated over the entire time period 2018/Q1-2025/Q3, five states account for over half (52.71%) of AI job postings in the U.S. economy: California (22.16%), Washington (9.26%), New York (8.32%), Texas (7.60%), and Massachusetts (5.37%). For California and New York, the high shares are a function of both their large size as well as above-average AI Jobs Intensity (California 0.81%, New York 0.72% - see Figure 5.2). Washington and Massachusetts are relatively smaller states but exhibit high levels of AI Jobs Intensity (Washington 1.40%, Massachusetts 0.67%); in fact, Washington state has the highest level of AI Jobs Intensity

across all geographies in the United States. In contrast, Texas is a large state but exhibits below average AI Jobs Intensity (0.34%).

As Figure 6.1 indicates, aggregated over the entire time period 2018/Q1-2025/Q3, one occupation alone (Computer and Mathematical, O*NET code 15) accounts for a dominant 69.12% share of AI-related job postings in the U.S. economy. This occupation accounts for a relatively small share of all job postings (7.06%, data not shown in the figure); however, the AI Jobs Intensity of this occupation is exceptionally high (3.80%, see Figure 6.2), almost 10 times the U.S. average of 0.39%. Other occupations at or above average levels of AI Jobs Intensity include Life, Physical, and Social Science (1.19%), Business and Financial Operations (0.87%), and Management (0.39%).

8. DISCUSSION AND CONCLUSIONS

This study advances research on AI adoption by tackling a foundational empirical challenge: how to measure firms' investment in AI-skilled human capital at scale and with precision. The dominant proxy used in prior strategy, management, economics, and finance research - AI-skilled job postings - has relied almost exclusively on the proprietary Lightcast-Burning Glass Skills Taxonomy. As documented in this paper, these legacy approaches systematically misclassify postings, often generating large numbers of false positives, leading to potentially unreliable conclusions about variations in AI adoption across firms, sectors, geographies, and occupations – cross-sectionally as well as longitudinally. In this paper, we develop an improved and transparent approach for identifying AI-related job postings and show that state-of-the-art LLMs, especially after task-specific fine-tuning, offer substantial gains in precision, computational efficiency, and scalability when benchmarked against the existing methods.

Our findings yield three major contributions. First, we demonstrate that modern LLMs, even at modest parameter scales, can outperform both closed-source frontier models and traditional LBG

taxonomy-based approaches on this core measurement task. We show that our fine-tuned Qwen 3-0.6B-DIS model outperforms all baseline and traditional methods, improving F1 by approximately 18% and precision by roughly 30% over taxonomy-based methods (see Table 6.2), with only 40% of the inference time as compared with that required by a leading larger-sized model (Qwen 3-8B-IT). These results suggest that domain-specific fine-tuning can close - as well as sometimes invert - the performance gap traditionally associated with model size. For strategy scholars, this insight underscores an important methodological implication: *carefully adapted smaller models can rival and even exceed the performance of much larger general-purpose LLMs*, offering an accessible path for large-scale, replicable text analysis.

Second, this paper introduces a transparent, high-performing pipeline for identifying AI jobs that is fully replicable and open-source. The combination of (a) a comprehensive dictionary-based prefilter, (b) a fine-tuned LLM classifier, and (c) a human-labeled ground-truth dataset provides a clear methodological roadmap for future work. Importantly, the improvements are not merely statistical but consequential: traditional skills-taxonomy based methods label more than half of supposed "AI jobs" incorrectly, distorting inferences about firm capability-building, labor demand, or competitive positioning. Our proposed approach directly addresses this issue by substantially reducing false positives without sacrificing recall.

Third, by applying our method to over seven years of LinkUp postings for Russell 1000 firms, we provide the first open-source dataset of quarterly AI-skilled labor demand for large U.S. firms from 2018 to 2025. This dataset can enable researchers to ask new questions about the diffusion of general-purpose technologies, the micro-foundations of digital strategy, and the antecedents and consequences of AI adoption. As described in the paper, AI hiring intensity in the U.S. economy rose from 0.22% in 2018 to 0.78% by 2025, with substantial heterogeneity across industries,

geographies, and occupations. These patterns not only illuminate where AI capability-building is most concentrated but also raise deeper questions about complementarities, organizational change, and the evolution of industry structure.

This study also provides broader lessons for scholars seeking to harness LLMs for empirical research. First, our results indicate that the marginal returns to model size diminish quickly; task-specific fine-tuning delivers larger performance gains than scaling up parameters. This insight complements emerging work using LLMs for measurement tasks and is consistent with the paper's discussion of the importance of domain adaptation for heterogeneous corpora such as job postings. Second, while few-shot prompting has attracted attention in computational linguistics, we show that few-shot prompting can worsen performance in highly heterogeneous business data settings by overly narrowing the model's decision boundaries. This suggests that scholars should be cautious when importing "best practices" from NLP benchmarks into applied research contexts.

More broadly, our study highlights the potential for frontier AI models to not only serve as a theoretical object of inquiry (i.e., how AI transforms organizations, work, and industries) but also to become powerful *methodological tools* that enable new forms of large-scale, fine-grained measurement. As strategy research increasingly turns to unstructured data (job postings, earnings calls, patent abstracts, product descriptions, employee reviews, and beyond), LLMs will become central to scaling qualitative insight into quantitative analysis. Our study demonstrates both the promise and the requirements for using LLMs responsibly in this role.

Like all data and measurement-focused studies, this paper has boundary limitations that open fruitful avenues for future work. First, although our dictionary plus LLM pipeline reduces both computational cost and false positives, dictionary prefilters necessarily rely on hand-built keyword lists that may require updating as AI-related terminology evolves. Second, while our LLM-based

classifier identifies whether a job requires *any* AI skill, future research may examine skill intensity at the level of specific AI skills, or complementarities between AI and non-AI skills. Third, future scholars could link our dataset to outcomes such as innovation, productivity, strategic renewal, performance heterogeneity, or market valuation, enabling richer causal inference regarding the returns to AI capability-building.

There is also an opportunity for cross-national comparative research. While our dataset focuses on U.S. firms, analogous pipelines could be applied to global job-posting datasets to study geographic diffusion of AI capabilities, competitive repositioning, or the emergence of new technological clusters worldwide.

Finally, as LLMs continue to evolve, the methodological frontier will continue shifting. Future work may explore multimodal job postings (text + images), agentic LLMs that autonomously retrieve additional information, or hybrid human-AI coding pipelines.

In sum, this study strengthens the empirical foundations of research on AI adoption by introducing a transparent, replicable, and highly accurate method for identifying AI-related jobs in online postings and by releasing the first longitudinal dataset of AI labor demand for large U.S. firms. By showing how fine-tuned LLMs dramatically improve measurement quality while reducing computational cost, this work highlights the value of domain-specific adaptation and sets the stage for a new line of strategy research that leverages frontier AI tools to illuminate how firms build the capabilities that shape competitive advantage in an AI-transformed economy.

ACKNOWLEDGEMENT

This research has been supported by an internal grant from the Smith School of Business, The University of Maryland.

References

Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy, 40*(121), 13-58.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics, 40*(S1), S293-S340.

Albert, D., Eklund, J. C., & Tang, L. (2025). A new organizational structure database: Examining structure through top management team compositions. *Strategic Management Journal,* Forthcoming. *https://sms.onlinelibrary.wiley.com/doi/pdf/10.1002/smj.70029*

Alekseeva, L., Azar, J., Giné, M., Samila, S., & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics, 71,* 102002.

Azar, J., Marinescu, I., Steinbaum, M., & Taska, B. (2020). Concentration in US labor markets: Evidence from online vacancy data. *Labour Economics, 66,* 101886.

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

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889-942.

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

Castilla, E.J., & Rho, H.J. (2023). The gendering of job postings in the online recruitment process. *Management Science, 69*(11), 6912-6939.

Chatterji, A., Cunningham, T., Deming, D.J., Hitzig, Z., Ong, C., Shan, C.Y., & Wadman, K. (2025). How people use ChatGPT. *NBER working paper 34255.*

Choi, J., Menon, A., & Tabakovic, H. (2021). Using machine learning to revisit the diversification–performance relationship. *Strategic Management Journal*, *42*, 1632–1661.

Choudhury, P., Wang, D., Carlson, N.A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, *40*, 1705–1732.

Choudhury, P., Allen, R.T., & Endres, M.G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*, 30–57.

Davis, J. (2021, December 20). The 2021 IT job market: What a year. *Information Week*. Retrieved from *https://www.informationweek.com/it-leadership/the-2021-it-job-market-what-a-year*

De Kok, T. (2025). ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Management Science*, *71*(9), 7888–7906.

Deming, D., & Kahn, L.B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics, 36*(S1). S337-S369.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., & Liu, T. (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Doshi, A.R., Bell, J.J., Mirzayev, E., & Vanneste, B.S. (2025). Generative artificial intelligence and evaluating strategic decisions. *Strategic Management Journal*, *46*, 583–610.

D'Acunto, F., Prabhala, N., & Rossi, A.G. (2019). The promises and pitfalls of robo-advising. *The Review of Financial Studies, 32*(5), 1983-2020.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science, 384*(6702), 1306-1308.

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

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5-47.

Goldfarb, A., He, X., & Teodoridis, F. (2025). Patterns of Artificial Intelligence Adoption by Hospitals. *AEA Papers and Proceedings, 115,* 40–45.

Goldfarb, A., Taska, B., & Teodoridis, F. (2020). Artificial intelligence in health care? Evidence from online job postings. *AEA Papers and Proceedings, 110,* 400-404.

Goldfarb, A., Taska, B., & Teodoridis, F. (2023). Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings. *Research Policy, 52*(1), 104653.

Gutiérrez, E., Lourie, B., Nekrasov, A., & Shevlin, T. (2020). Are online job postings informative to investors? *Management Science, 66*(7), 3133-3141.

Ham, C., Hann, R.N., Rabier, M., & Wang, W. (2025). Auditor skill demands and audit quality: Evidence from job postings. *Management Science, 71*(7), 5805-5829.

Hampole, M., Papanikolaou, D., Schmidt, L.D., & Seegmiller, B. (2025). Artificial intelligence and the labor market, Working Paper. Available at SSRN: *https://ssrn.com/abstract=5121025*.

Harrison, J.S., Josefy, M.A., Kalm, M., & Krause, R. (2023). Using supervised machine learning to scale human-coded data: A method and dataset in the board leadership context. *Strategic Management Journal*, *44*, 1780–1802.

Hershbein, B., & Kahn, L.B. (2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review, 108*(7), 1737-1772.

Huang, A.H., Pacelli, J., Shi, T.T., & Zou, Y. (2024). Communicating corporate culture in labor markets: Evidence from job postings. Working Paper. Available at SSRN: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4235342* .

Kim, A.G., Muhn, M., Nikolaev, V.V., & Zhang, Y. (2025). Learning fundamentals from text. Chicago Booth Accounting Research Center. Working Paper. Available at. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5047947#*

Lee, S., Kim, J.D. (2024). When do startups scale? Large-scale evidence from job postings. *Strategic Management Journal,* 45(9), 1633-1669.

Luo, X., Jia, N., & Ouyang, E., & Fang, Z. 2024. Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal*, *45*, 1597–1629.

Mah, P. (2021 November 3). How Apple uses AI To produce better photos. Retrieved from *https://www.cdotrends.com/story/15984/how-apple-uses-ai-produce-better-photos*.

Makridis, C., & Alterovitz, G. (2024). Measuring and Understanding Differences in Private and Public Sector Technology Jobs: Evidence from Artificial Intelligence Job Posting Data. Working Paper. Available at SSRN: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4891300*

Meister, D., & Chandrasekhar, R. (2021). McCormick & Co.: Deploying Artificial Intelligence in New Product Development. Case Study W25509. Ivy Publishing, Available at https://myscma.com/wp-content/uploads/2024/03/McCormick-Co-Case-Study.pdf.

Miric, M., Jia, N., & Huang, K.G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, *44*, 491–519.

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

Oremus, W. (2015). What is TensorFlow, and why is Google so excited about it? *Slate. http://www.slate.com/blogs/future_tense/2015/11/09/google_s_tensorflow_is_open_source_and_it_s_about_to_be_a_huge_huge_deal. html*.

Otis, N., Clarke, R., Delecourt, S., Holtz, D., & Koning, R. (2024). The uneven impact of generative AI on entrepreneurial performance. UC-Berkeley Haas School of Business, Working Paper.

Rathje, J., Katila, R., & Reineke, P. (2024). Making the most of AI and machine learning in organizations and strategy research: Supervised machine learning, causal inference, and matching models. *Strategic Management Journal, 45*, 1926–1953.

Shao, Y., Zope, H., Jiang, Y., Pei, J., Nguyen, D., Brynjolfsson, E., & Yang, D. (2025). Future of work with AI agents: Auditing automation and augmentation potential across the US workforce. arXiv preprint *arXiv:2506.06576*.

Squicciarini, M., & Nachtigall, H. (2021). Demand for AI skills in jobs: Evidence from online job postings. *OECD Science, Technology and Industry Working Papers, 2021*(3), 1-74.

Stanford University. (2025). The 2025 AI Index Report. Available at *https://hai.stanford.edu/ai-index/2025-ai-index-report.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*, 1-11.

Von Krogh, G., Roberson, Q., & Gruber, M. (2023). From the Editors - Recognizing and utilizing novel research opportunities with artificial intelligence. *Academy of Management Journal*, *66*(2), 367-373.

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A.M., & Le, Q.V. (2021). Finetuned language models are zero-shot learners. arXiv preprint *arXiv:2109.01652*.

West, D.M., & Allen, J.R. (2018). How artificial intelligence is transforming the world. Brookings. *https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/*

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A.M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 EMNLP (Systems Demonstrations)*, 38–45.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., …. Qiu, Z. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388.*

Zhang, B., Liu, Z., Cherry, C., & Firat, O. (2024). When scaling meets llm finetuning: The effect of data, model and finetuning method. arXiv preprint *arXiv:2402.17193*.

TABLE 1

How Different Studies Have Used LBG Skills Taxonomy to Identify AI Jobs

Study	**Data Source for Classification**	**Specific Methodology for Identifying an AI Job in Job Postings**
Goldfarb et al. (2020)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	Any job that is labeled by LBG-ST as belonging to one of the following broad skill clusters: "Artificial Intelligence," "Machine Learning," and "Natural Language Processing."
Alekseeva et. (2021)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	Any job that is labeled by LBG-ST as requiring any of 71 specific AI skills.
Squicciarini and Nachtigall (2021)	Glass Skills Taxonomy (LBG-ST)	Any job that contains at least two AI-related keywords in the LBG-ST list of skills (one of them need not be software-related).
Acemoglu et al. (2022)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	Any job that is labeled by LBG-ST as requiring any of 33 specific AI skills.
Goldfarb et al. (2023)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	Any job that is labeled by LBG-ST as requiring at least one skill in the broad skill cluster of "Machine Learning."
Makridis and Alterovitz (2024)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	Any job that is labeled by LBG-ST as requiring any of 172 specific AI skills.
Babina et al. (2024)	Lightcast-Burning Glass Skills Taxonomy (LBG-ST)	*Step 1:* Using the entire database in LBG-ST, compute the *AI-relatedness* of each skill by calculating its co-occurrence with the following skills: "Artificial Intelligence (AI)," "Machine Learning (ML)," "Natural Language Processing (NLP)," and "Computer Vision (CV)." *Step 2:* For each job posting, compute the mean AI-relatedness across all skills associated with that job. *Step 3:* Treat the job posting as an AI job if the mean AI-relatedness across all skills associated with that job (as computed in Step 2) is above 0.1.

TABLE 2

Performance Comparison Across LLMs Within and Across LLM-Families (Zero-Shot Classification)

N = 200,000 (Ground Truth Dataset)
Two-Step Evaluation Method: Keyword Filtering Followed by LLM-Based Classification[a]
Mean Values (and Standard Deviations) Across Five Runs

	F1	Precision	Recall	Accuracy
Closed-Source Models				
OpenAI GPT-4o	0.762 (0.003)	0.773 (0.004)	0.751 (0.002)	0.995 (0.000)
Anthropic Claude-3.7	0.738 (0.001)	0.788 (0.002)	0.695 (0.002)	0.994 (0.000)
Open-Source Models				
Llama Family				
Llama 3.2-3B-Instruct	0.342 (0.037)	0.956 (0.005)	0.209 (0.028)	0.991 (0.000)
Llama 3.2-1B-Instruct	0.348 (0.005)	0.246 (0.002)	0.596 (0.023)	0.974 (0.001)
Llama 3.1-8B-Instruct	0.751 (0.003)	0.654 (0.003)	0.880 (0.007)	0.993 (0.000)
Qwen Family				
Qwen 3-14B	0.791 (0.002)	0.811 (0.004)	0.772 (0.008)	0.995 (0.000)
Qwen 3-8B	0.764 (0.002)	0.668 (0.006)	0.892 (0.007)	0.994 (0.000)
Qwen 3-4B	0.690 (0.007)	0.545 (0.010)	0.939 (0.006)	0.990 (0.000)
Qwen 3-1.7B	0.571 (0.028)	0.426 (0.037)	0.871 (0.040)	0.985 (0.000)
Qwen 3-0.6B	0.075 (0.033)	0.360 (0.042)	0.042 (0.020)	0.988 (0.003)
Qwen 2.5-14B-Instruct	0.726 (0.010)	0.878 (0.005)	0.619 (0.017)	0.995 (0.000)
Qwen 2.5-7B-Instruct	0.726 (0.006)	0.771 (0.011)	0.686 (0.019)	0.994 (0.000)
Mistral Family				
Mistral v0.3-7B-Instruct	0.752 (0.004)	0.722 (0.008)	0.785 (0.001)	0.994 (0.000)
Mistral Nemo-Instruct	0.746 (0.004)	0.661 (0.012)	0.858 (0.011)	0.993 (0.000)

[a]Following the rationale in Sections 4.2, we partition the dataset using a comprehensive dictionary of AI-related keywords (Appendix 2) into postings with and without such terms. As the discussion in Section 4.2 also clarifies, postings lacking any AI-related keyword can be robustly classified as Truly-Not-AI, while those containing at least one AI-related keyword may include both Truly-AI and Truly-Not-AI jobs. We thus apply automated classification methods only to this latter subset. This two-step approach preserves classification accuracy while reducing computational and financial costs by roughly 95%, given that keyword-positive postings constitute only about 5% of the total dataset.

TABLE 3

Performance Comparison Across LLMs Within and Across LLM-Families (Few-Shot Classification and Improvement Over Zero-Shot Classification)

N = 200,000 (Ground Truth Dataset)

Two-Step Evaluation Method: Keyword Filtering Followed by LLM-based Classification[b]

Mean Values (and Standard Deviations) Across Five Runs

	F1	Precision	Recall	Accuracy	ΔF1	ΔPrec	ΔRec	ΔAcc
Closed-Source Models								
OpenAI GPT-4o	0.789 (0.001)	0.792 (0.003)	0.786 (0.004)	0.995 (0.000)	0.027	0.019	0.035	0.001
Anthropic Claude-3.7	0.623 (0.001)	0.773 (0.001)	0.521 (0.002)	0.993 (0.000)	-0.115	-0.015	-0.174	-0.002
Open-Source Models								
Llama Family								
Lamma 3.2-3B-Instruct	0.012 (0.002)	0.987 (0.030)	0.006 (0.001)	0.988 (0.000)	-0.329	0.031	-0.202	-0.002
Lamma 3.2-1B-Instruct	0.369 (0.004)	0.241 (0.002)	0.789 (0.018)	0.968 (0.000)	0.021	-0.007	0.193	-0.005
Lamma 3.1-8B-Instruct	0.762 (0.004)	0.712 (0.004)	0.820 (0.004)	0.994 (0.000)	0.012	0.058	-0.06	0.001
Qwen Family								
Qwen 3-14B	0.772 (0.001)	0.849 (0.001)	0.708 (0.002)	0.995 (0.000)	-0.019	0.038	-0.064	0.000
Qwen 3-8B	0.776 (0.001)	0.734 (0.001)	0.823 (0.001)	0.994 (0.000)	0.012	0.066	-0.069	0.001
Qwen 3-4B	0.710 (0.000)	0.587 (0.000)	0.896 (0.001)	0.991 (0.000)	0.021	0.042	-0.043	0.001
Qwen 3-1.7B	0.659 (0.001)	0.597 (0.002)	0.736 (0.002)	0.991 (0.000)	0.089	0.171	-0.135	0.007
Qwen 3-0.6B	0.331 (0.005)	0.255 (0.004)	0.472 (0.007)	0.978 (0.000)	0.261	-0.105	0.430	-0.010
Qwen 2.5-14B-Instruct	0.649 (0.001)	0.945 (0.000)	0.494 (0.001)	0.994 (0.000)	-0.077	0.068	-0.124	-0.001
Qwen 2.5-7B-Instruct	0.678 (0.001)	0.812 (0.001)	0.582 (0.002)	0.994 (0.000)	-0.048	0.042	-0.104	0.000
Mistral Family								
Mistral v0.3-7B-Instruct	0.767 (0.000)	0.800 (0.000)	0.737 (0.000)	0.994 (0.000)	0.015	0.078	-0.048	0.000
Mistral Nemo-Instruct	0.671 (0.000)	0.829 (0.000)	0.564 (0.000)	0.994 (0.000)	-0.075	0.168	-0.294	0.001

[b] Following the rationale in Section 4.2, we partition the dataset using a comprehensive dictionary of AI-related keywords (Appendix 2) into postings with and without such terms. As the discussion in Section 4.2 also clarifies, postings lacking any AI-related keyword can be robustly classified as Truly-Not-AI, while those containing at least one AI-related keyword may include both Truly-AI and Truly-Not-AI jobs. We thus apply automated classification methods only to this latter subset. This two-step approach preserves classification accuracy while reducing computational and financial costs by roughly 95%, given that keyword-positive postings constitute only about 5% of the total dataset.

TABLE 4

Performance Comparison Across Supervised Fine Tuning Open-Source LLMs

N = 40,000 (Ground Truth Dataset)
Two-Step Evaluation Method: Keyword Filtering Followed by LLM-Based Classification[c]

Results from Two Supervised Fine-Tuning Approaches Applied to the Two Best-Performing Open-Source Models (Qwen 3-14B and Qwen 3-8B, Based on F1 Score) and the Most Cost-Efficient Model (Qwen 3-0.6B)

Supervised Fine-Tuned Model	Performance Metrics				Improvement Over Models Without Fine Tuning			
	F1	**Precision**	**Recall**	**Accuracy**	**ΔF1**	**ΔPrec**	**ΔRec**	**ΔAcc**
Zero-shot After Instruction Tuning								
Qwen 3-14B-IT	0.817	0.806	0.828	0.997	0.042	-0.003	0.085	0.002
Qwen 3-8B-IT	0.831	0.820	0.843	0.997	0.075	0.160	-0.040	0.003
Qwen 3-0.6B-IT	0.806	0.872	0.749	0.996	0.730	0.506	0.706	0.008
Few-shot After Instruction Tuning								
Qwen 3-14B-IT	0.795	0.792	0.799	0.996	0.049	-0.046	0.126	0.001
Qwen 3-8B-IT	0.803	0.817	0.790	0.997	0.032	0.081	-0.019	0.002
Qwen 3-0.6B-IT	0.797	0.832	0.765	0.996	0.474	0.586	0.297	0.018
Fine-Tuned Discriminator								
Qwen 3-14B-DIS	0.820	0.836	0.803	0.996				
Qwen 3-8B-DIS	0.827	0.871	0.787	0.996				
Qwen 3-0.6B-DIS	0.827	0.852	0.803	0.996				

[c] Recall that the overall ground-truth dataset consists of 200,000 job postings. Following the standard 80-20 train-test split widely used in machine learning, we used 80% of the dataset for fine-tuning and 20% for the evaluations reported in this table. Consistent with the logic detailed in the footnotes to Tables 2 and 3, we adopt a 2-step approach for the results presented here. We first apply a comprehensive dictionary of AI-related keywords and then apply the automated method to the subset with at least one AI-related keyword.

Note: "IT" refers to instruction tuning, a fine-tuning setup in which the model retains its conversational and generative capabilities. In contrast, "DIS" denotes the discriminator setup, where the model is structurally modified into a binary classifier. Under this configuration, the model no longer functions as a conversational agent. Accordingly, for the discriminator models, we input the raw job description directly for classification, without providing any task-specific instruction (as in zero-shot classification) or example demonstrations (as in few-shot classification). For the discriminator models, we do not use the base models' zero-shot or few-shot results to measure improvement, because these approaches represent fundamentally different paradigms – the former being generative and the latter discriminative in nature.

TABLE 5

Comparison of Computing Time and Monetary Cost Across Models

	Zero-shot		Few-shot	
	Time (s)	**Cost ($)**	**Time (s)**	**Cost ($)**
Closed-Source Models				
OpenAI GPT-4o	4,340	1.60	5,161	4.33
Anthropic Claude-3.7	24,548	2.17	25,437	5.87
Open-Source Models without Fine-Tuning				
Llama Family				
Lamma 3.2-3B-Instruct	751	0.13	1,128	0.19
Lamma 3.2-1B-Instruct	441	0.07	578	0.10
Lamma 3.1-8B-Instruct	1,062	0.18	1912	0.32
Qwen Family				
Qwen 3-14B	1,706	0.28	3,319	0.55
Qwen 3-8B	1,275	0.21	2,183	0.36
Qwen 3-4B	1,143	0.19	1,647	0.28
Qwen 3-1.7B	848	0.14	993	0.17
Qwen 3-0.6B	841	0.14	919	0.15
Qwen 2.5-14B-Instruct	1,692	0.28	3,266	0.54
Qwen 2.5-7B-Instruct	937	0.16	1,700	0.28
Mistral Family				
Mistral v0.3-7B-Instruct	1,066	0.18	2,044	0.34
Mistral Nemo-Instruct	1,402	0.23	2,729	0.46
Open-Source Models with Fine-Tuning				
Models with Instruction Tuning				
Qwen 3-14-IT	2,959	0.48	4,733	0.78
Qwen 3-8B-IT	2,182	0.36	3,198	0.54
Qwen 3-0.6B-IT	1,513	0.25	1,578	0.25
Discriminator Models				
Qwen 3-14B-DIS	2,559	0.42	-	-
Qwen 3-8B-DIS	1,566	0.25	-	-
Qwen 3-0.6B-DIS	855	0.15	-	-

Notes: (1) Reported costs are based on processing 10,000 jobs and the costs are presented in both seconds of compute time and U.S. dollars. (2) The values reported for the discriminator models under the "Zero-shot" columns reflect the cost specific to the discriminative classification setup, rather than a conversational generative setting.

TABLE 6.1

Benchmarking Top-Performing LLM Method Against Existing Methods

N = 40,000 (Ground Truth Dataset)

2-Step Method for Evaluations: Keywords Filtering Followed by Automated Method[d]

Methods	F1	Precision	Recall	Accuracy
LBG Skill Label (Acemoglu *et al.* 2022)	0.684	0.590	0.813	0.992
LBG Skill-Relatedness (Babina *et al.*, 2024)	0.657	0.535	0.852	0.990
Top-Performing LLM Method: Qwen 3-0.6B-DIS	0.827	0.852	0.803	0.996

[d]Recall that the overall ground-truth dataset consists of 200,000 job postings. Following the standard 80-20 train-test split widely used in machine learning, we used 80% of the dataset for fine-tuning and 20% for the evaluations reported in this table. Consistent with the logic detailed in the footnotes to Tables 2 and 3, we adopt a 2-step approach for the results presented here. We first apply a comprehensive dictionary of AI-related keywords and then apply the automated method to the subset with at least one AI-related keyword. For the sake of consistency, we implement this 2-step approach for the LBG Skill Label (Acemoglu et al., 2022) and the LBG Skill-Relatedness (Babina et al., 2024) methods as well.

TABLE 6.2

Benchmarking Top-Performing LLM Method Against Existing Methods

N = 40,000 (Ground Truth Dataset)

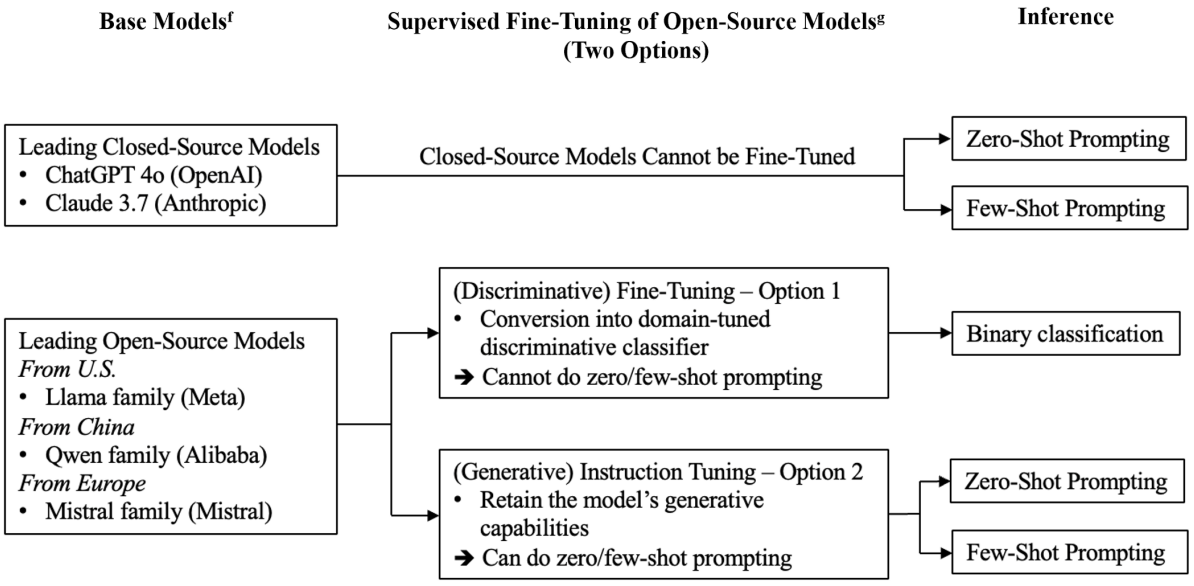
No Keywords Filtering, Direct Application of Automated Method Only[e]

Methods:	F1	Precision	Recall	Accuracy
LBG Skill Label (Acemoglu *et al.* 2022)	0.610	0.489	0.813	0.988
LBG Skill-Relatedness (Babina *et al.*, 2024)	0.587	0.448	0.852	0.987
Top-Performing LLM Method: Qwen 3-0.6B-DIS	0.791	0.780	0.803	0.995

[e]In this table, we skip the first step used in Table 6.1. Specifically, we do not apply any AI-related keyword filtering to the 40,000 job postings in the test set. We take this approach because the LBG Skill Label (Acemoglu et al., 2022) and the LBG Skill-Relatedness (Babina et al., 2024) methods do not employ the 2-step approach that we have argued earlier to be more efficient. By skipping the filtering step, we ensure that our comparison remains methodologically consistent with the original implementations of the LBG approaches. This design allows for a more direct comparison between the top-performing LLM-based method (Qwen 3-0.6B-DIS) and the LBG methods as originally proposed by Acemoglu et al. (2022) and Babina et al. (2024).

Figure 1

Approach to Evaluation of LLM-Based Methods

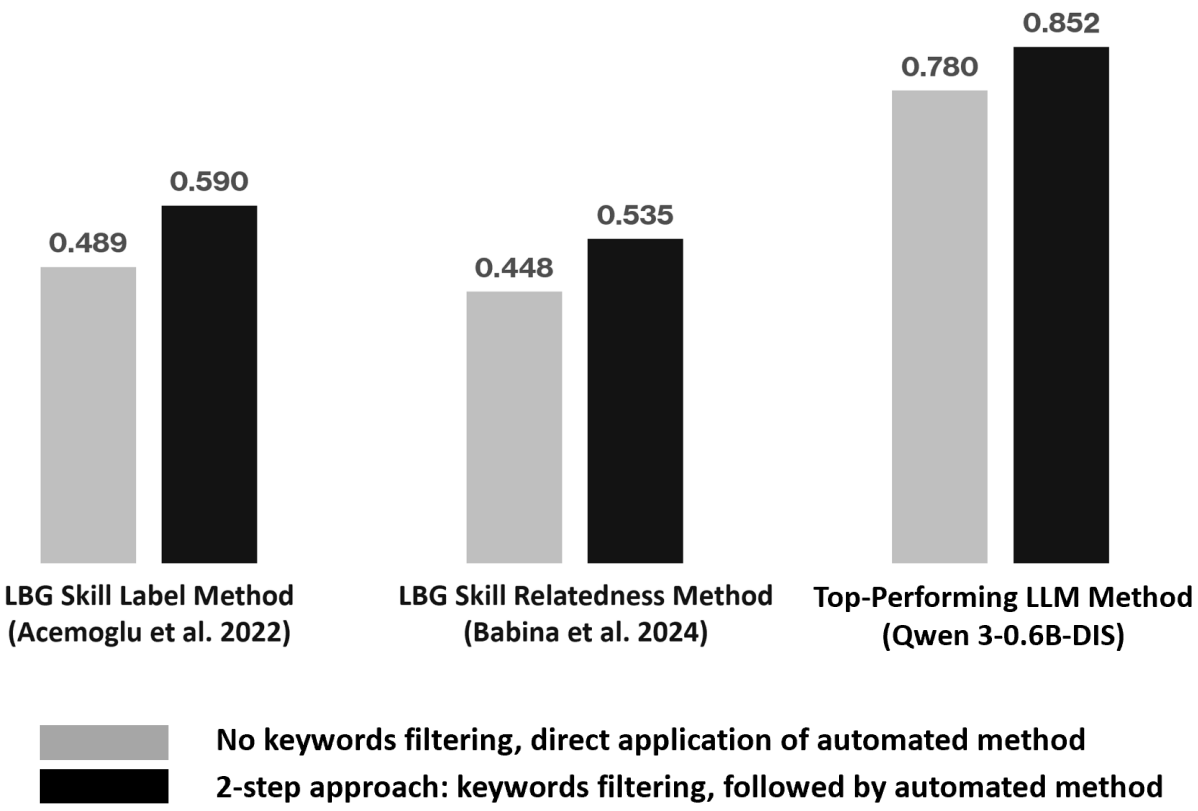


[f]Leading base models that were publicly available as of June 2025
[g]Conducted only the two best-performing open-source models (based on F1 Score) and the most cost-efficient model, before fine-tuning

FIGURE 2

Relative Precision of Traditional Methods and Top-Performing LLM Method

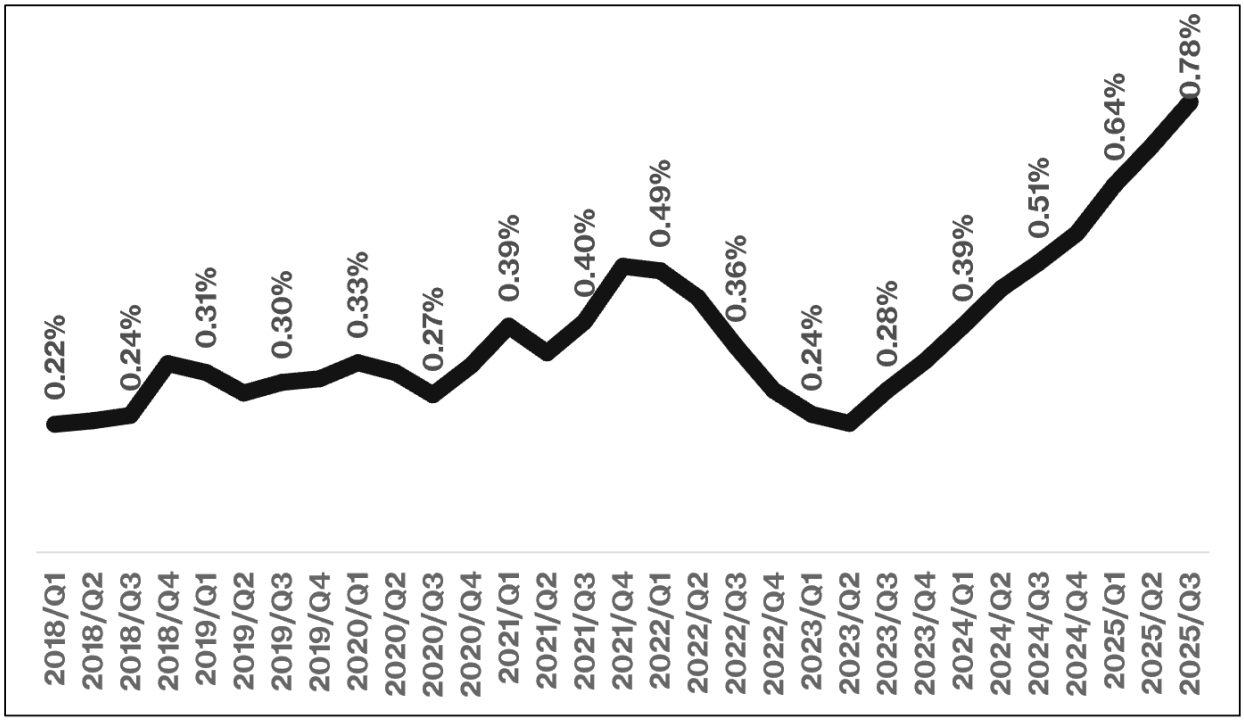
Precision = A measure of the accuracy of positive predictions i.e., when the model says "yes," how often it is correct = True Positives/(True Positives + False Positives)



Notes: The figure demonstrates that the two-step approach increases the performance of every automated classification method. Importantly, as detailed in the footnotes to Table 2, this procedure yields substantial efficiency gains. By first filtering out clearly irrelevant postings through a keyword-based pre-screen and then applying the LLM-based classifier to the remaining subset, the two-step approach reduces computational and monetary cost by approximately 95%, while preserving classification accuracy.

FIGURE 3

AI Jobs Intensity of the U.S. Economy



Notes: (1) These data come from the application of the top-performing LLM method to the universe of all job postings in the U.S. economy (raw data source: LinkUp). (2) AI Jobs Intensity is the ratio of the number of job postings for AI-related jobs to the number of job postings for all jobs.

FIGURE 4.1

% Share of AI-Related Job Postings in the U.S. Economy – Across Sectors
(2018/Q1 Through 2025/Q3, Cumulative)

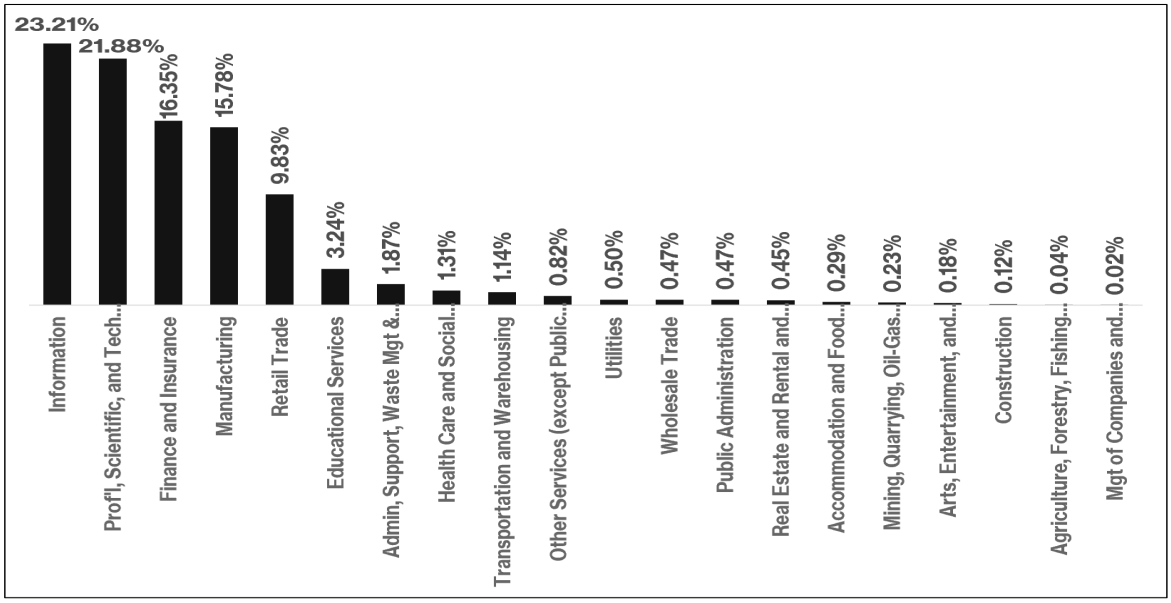
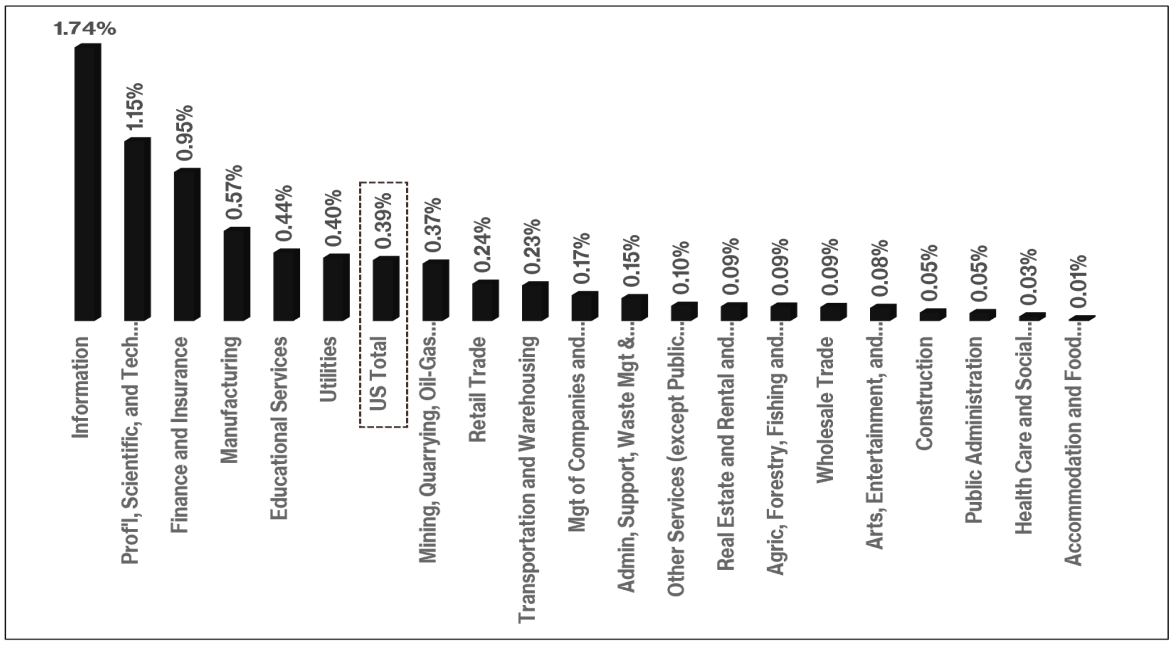


FIGURE 4.2

AI Jobs Intensity of the U.S. Economy – Across Sectors
(2018/Q1 Through 2025/Q3, Cumulative)



Notes: (1) These data come from the application of the top-performing LLM method to the universe of all job postings in the U.S. economy (raw data source: LinkUp). (2) Sectors refer to 2-digit NAICS sectors. (3) AI Jobs Intensity is the ratio of the number of job postings for AI-related jobs to the number of job postings for all jobs.

FIGURE 5.1

% Share of AI-Related Job Postings in the U.S. Economy – Across States (2018/Q1 Through 2025/Q3, Cumulative)

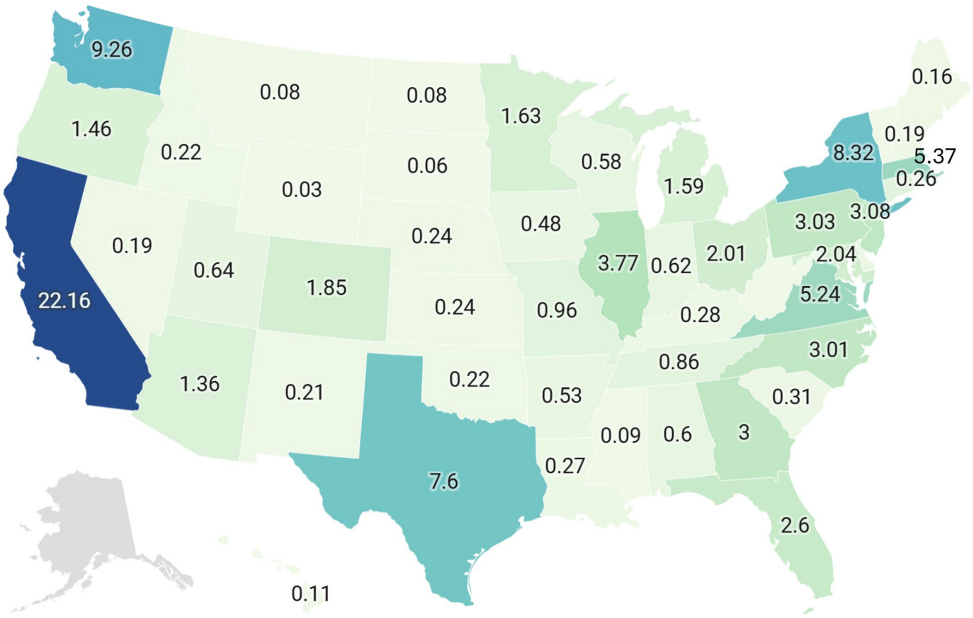
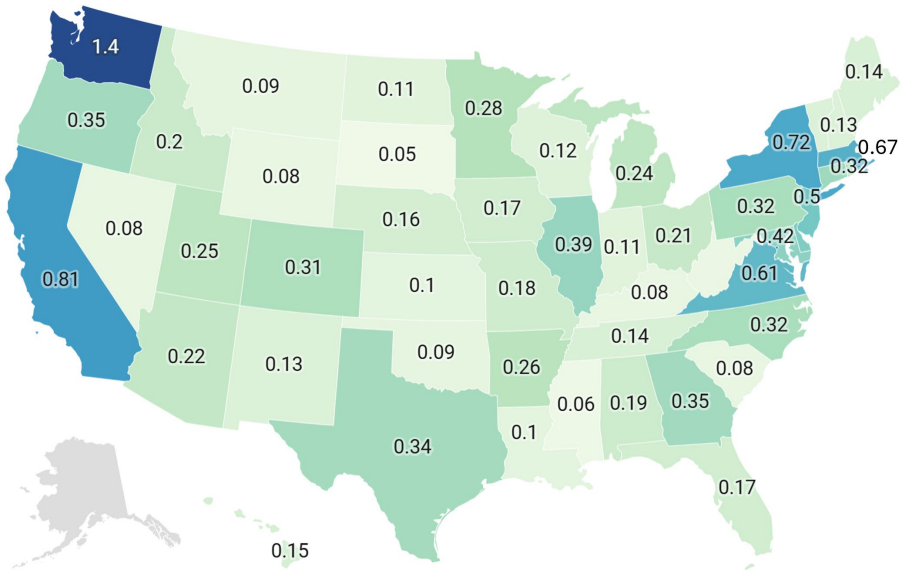


FIGURE 5.2

AI Jobs Intensity of the U.S. Economy – Across States (2018/Q1 Through 2025/Q3, Cumulative)



Notes: (1) These data come from the application of the top-performing LLM method to the universe of all job postings in the U.S. economy (raw data source: LinkUp). (2) AI Jobs Intensity is the ratio of the number of job postings for AI-related jobs to the number of job postings for all jobs.

FIGURE 6.1

% Share of AI-Related Job Postings in the U.S. Economy – Across Occupations (2018/Q1 Through 2025/Q3, Cumulative)

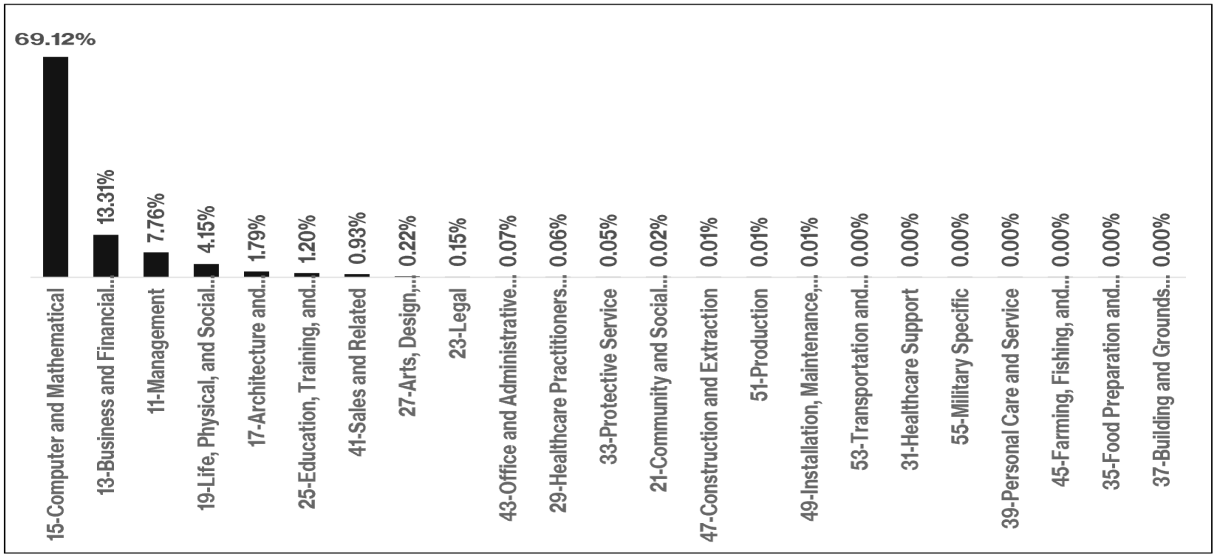
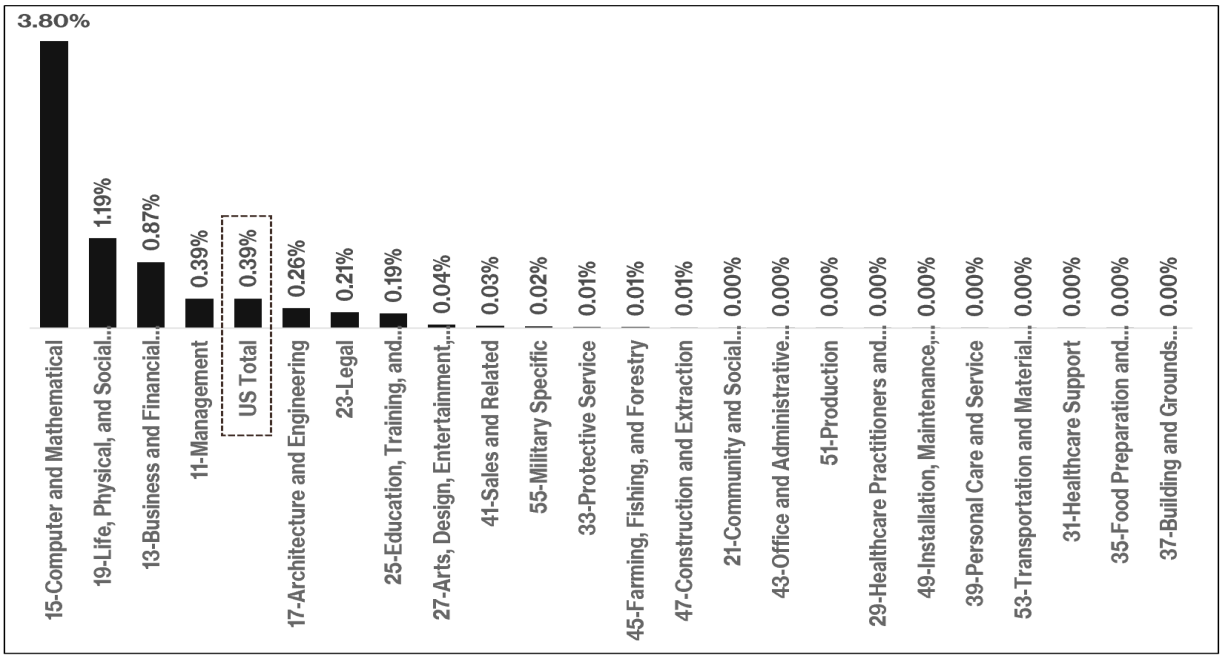


FIGURE 6.2

AI Jobs Intensity of the U.S. Economy – Across Occupations (2018/Q1 Through 2025/Q3, Cumulative)



Notes: (1) These data come from the application of the top-performing LLM method to the universe of all job postings in the U.S. economy (raw data source: LinkUp). (2) Occupations refer to 2-digit O*NET occupations. (3) AI Jobs Intensity is the ratio of the number of job postings for AI-related jobs to the number of job postings for all jobs.

APPENDIX 1

Zero-shot and Few-shot Prompt Templates

Zero-Shot Prompt Template:

Carefully read the following job description and determine whether the role requires artificial intelligence (AI) related technical skills, such as but not limited to neural networks, large language models, computer vision, deep learning, pre-training / fine-tuning LLMs, and similar technologies. If yes, label it as 1 for an AI job and 0 otherwise. Keep in mind that merely knowing how to "use" an AI tool (e.g., ChatGPT or Llama) does not make it an AI-skilled job. Consider an analogy. Knowing how to use Microsoft PowerPoint does not require any technical skill; however, developing or upgrading the PowerPoint software does require technical skills.

Strictly respond with a single number only: 1 or 0 (1 for AI job and 0 for Non-AI jobs). No reasoning or explanation needed.

Few-Shot Prompt Template:

Carefully read the following job description and determine whether the role requires artificial intelligence (AI) related technical skills, such as but not limited to neural networks, large language models, computer vision, deep learning, pre-training / fine-tuning LLMs, and similar technologies. If yes, label it as 1 for an AI job and 0 otherwise. Keep in mind that merely knowing how to "use" an AI tool (e.g., ChatGPT or Llama) does not make it an AI-skilled job. Consider an analogy. Knowing how to use Microsoft PowerPoint does not require any technical skill; however, developing or upgrading the PowerPoint software does require technical skills.

To understand better, here are examples of AI Jobs (label = 1): {AI-Job Examples} and here are non-AI job examples (label = 0): {Non-AI Job Examples}

Strictly respond with a single number only: 1 or 0 (1 for AI job and 0 for Non-AI jobs). No reasoning or explanation needed.

APPENDIX 2

AI-Skill Keyword Dictionary Used to Identify Seemingly-AI Jobs

ai, artificial intelligence, ml, machine learning, deep learning, neural networks, supervised learning, unsupervised learning, reinforcement learning, transfer learning, semi-supervised learning, ensemble learning, feature engineering, dimensionality reduction, decision trees, random forests, svm, support vector machines, naive bayes, k-nearest neighbors, knn, instance-based learning, model selection, model evaluation, hyperparameter optimization, overfitting, underfitting, cross-validation, bias-variance tradeoff, gradient descent, backpropagation, loss function, activation function, optimization algorithms, stochastic gradient descent, sgd, batch gradient descent, mini-batch gradient descent, adam optimization, batch normalization, data augmentation, recurrent neural networks, rnn, long short-term memory, lstm, gated recurrent unit, gru, convolutional neural networks, cnn, convolutional layers, pooling layers, fully connected layers, object detection, image segmentation, generative adversarial networks, gans, autoencoders, natural language processing, nlp, word embeddings, named entity recognition, sentiment analysis, text classification, text generation, seq2seq models, attention mechanism, transformers, pre-trained models, computer vision, image classification, image recognition, image captioning, optical character recognition, ocr, time series analysis, anomaly detection, recommender systems, collaborative filtering, reinforcement learning algorithms, policy gradient methods, markov decision processes, monte carlo methods, alphago, alphazero, robotics, internet of things, data science, big data, data mining, pattern recognition, explainable ai, bayesian networks, genetic algorithms, evolutionary computation, swarm intelligence, knowledge representation, expert systems, fuzzy logic, cognitive computing, edge computing, cloud computing, large language models, llms, generative models, computational creativity, image synthesis, text generation, algorithm development, algorithmic optimization, probabilistic models, computer graphics, augmented reality, virtual reality

APPENDIX 3

Annotation Instructions

The following instruction was provided to the 10 annotators:

Carefully read the following job description and determine whether the role requires artificial intelligence (AI) related technical skills, such as but not limited to neural networks, large language models, computer vision, deep learning, pre-training / fine-tuning LLMs, and similar technologies. If yes, label it as 1 for an AI job and 0 otherwise. Keep in mind that merely knowing how to "use" an AI tool (e.g., ChatGPT or Llama) does not make it an AI-skilled job. Consider an analogy. Knowing how to use Microsoft PowerPoint does not require any technical skill; however, developing or upgrading the PowerPoint software does require technical skills.

APPENDIX 4

Consistency Analysis of LLM-based Methods

Another important consideration for business research using LLMs is consistency, as it directly affects reproducibility. To assess this dimension, we report the standard deviation across five independent runs, presented in parentheses below each mean value in Table 2 and Table 3. Generative LLMs inherently contain randomness. Because these models generate text token-by-token, their outputs depend on probabilistic sampling strategies governed by hyperparameters such as temperature, top-*k*, and top-*p*. While one could eliminate randomness by employing greedy decoding (e.g., setting temperature = 0), prior technical reports caution that this approach often leads to degraded performance and is not recommended by model developers (Yang et al., 2025).

In contrast, skill-taxonomy-based methods exhibit much higher consistency by producing deterministic classification labels rather than free-form text. Both LBG Skill Label and the LBG Skill Relatedness methods achieve perfect reproducibility, with zero standard deviation across runs. Similarly, discriminatively tuned LLM approaches (e. g., Qwen 3-0.6B-DIS) directly transform text into binary classification labels, thereby circumventing the randomness inherent in text generation and producing highly stable results.

For generative LLMs reported in Table 2, model size plays an important role in consistency. Smaller models (fewer than 4 billion parameters) show greater variability across runs, whereas larger models (7 billion parameters and above) achieve very low variation on F1 scores (typically <0.01). This pattern suggests that researchers seeking reproducible results should either use sufficiently large generative models or opt for non-generative approaches when computational constraints or inference stability are of concern. Notably, closed source LLMs also show comparably low variability.

APPENDIX 5

Threshold Sensitivity Analysis of the LBG Skill-Relatedness Method (Babina et al., 2024)

As the following sensitivity analyses show, irrespective of whether one uses the 2-step approach (with keywords filtering) or the 1-step approach (without any keywords filtering), based on F1 scores, the performance of the LBG Skill-Relatedness method degrades rapidly at higher thresholds. While precision improves, recall drops sharply. We conjecture that the rapid drop in recall (i.e., failing to classify "true AI jobs" as "predicted AI jobs") happens because of two major weaknesses of this method that we discussed on earlier in Section 2.2: (1) the lower-than-valid AI-relatedness score attributed to many AI-specific skills (e.g., "Supervised Learning (Machine Learning)" has an AI-relatedness of only 0.695 and the skill "Neural Networks" has an AI-relatedness of only 0.640 respectively, even though both skills have a definitive AI-relatedness of 1.0); (2) assignment of equal weight to every skill – whether "Neural Networks" or "Leadership" - when computing the mean AI-relatedness across all skills associated with that job.

Table A1.1

N = 40,000 (Ground Truth Dataset)

2-Step Method for Evaluations: Keywords Filtering Followed by Automated Method

Performance of LBG Skill-Relatedness Method				
Threshold Above Which AI-Relatedness is Used to Classify a Job as an AI Job	**F1**	**Precision**	**Recall**	**Accuracy**
0.1	0.657	0.535	0.852	0.990
0.2	0.606	0.814	0.482	0.993
0.3	0.344	0.896	0.213	0.991
0.4	0.133	0.889	0.072	0.990
0.5	0.044	1.000	0.022	0.989
0.6	0.013	1.000	0.007	0.989
0.7	0.004	1.000	0.002	0.989
0.8	0	0	0	0.989
0.9	0	0	0	0.989

Table A1.2

N = 40,000 (Ground Truth Dataset)

No Keywords Filtering, Direct Application of Automated Method Only

Performance of LBG Skill-Relatedness Method				
Threshold Above Which AI-Relatedness is Used to Classify a Job as an AI Job	**F1**	**Precision**	**Recall**	**Accuracy**
0.1	0.587	0.448	0.852	0.986
0.2	0.592	0.768	0.482	0.993
0.3	0.343	0.880	0.213	0.991
0.4	0.132	0.842	0.072	0.990
0.5	0.044	1.000	0.022	0.989
0.6	0.013	1.000	0.007	0.989
0.7	0.004	1.000	0.002	0.989
0.8	0	0	0	0.989
0.9	0	0	0	0.989