RESEARCH REPORT
STUDY PAPER 281
MARCH 2026

LLMs as Economic Measurement Tools: A Framework for Multi-Label Job Posting Classification

Rentian Zhu and Daniel Hardt

THE ROCKWOOL FOUNDATION
RESEARCH

STUDY PAPER 281 MARCH 2026

LLMs as Economic Measurement Tools: A Framework for Multi-Label Job Posting Classification

Published by:

Address:
The ROCKWOOL Foundation Research Unit
Ny Kongensgade 6
1472 Copenhagen, Denmark

Telephone +45 33 34 48 00
E-mail: kontakt@rff.dk
en.rockwoolfonden.dk/research/

March 2026

LLMs as Economic Measurement Tools: A Framework for Multi-Label Job Posting Classification

Rentian Zhu* Daniel Hardt †

November 14, 2025

Abstract

This paper introduces large language models (LLMs) as measurement instruments for extracting economic variables from unstructured text. Using 986 manually-annotated Danish job postings, we develop and validate a human-in-the-loop framework that constructs a 25-category classification system capturing skill requirements and workplace flexibility dimensions. Our analysis yields four key findings. First, LLMs substantially outperform traditional keyword-based methods, with an increase in accuracy from 73.4% to 85.3%, for base models, and 93% accuracy for fine-tuned models. Second, we show that open-source models (Llama-3.3-70B) can achieve performance parity with OpenAI's GPT models while permitting on-premise deployment, thereby addressing critical data confidentiality constraints for administrative and confidential microdata that have limited LLM adoption in economic research. Third, we document an inverse relationship between model confidence (log probabilities) and classification errors, enabling researchers to achieve higher accuracy through confidence-based sample selection without additional training costs. Fourth, we demonstrate the economic relevance of these measures by classifying 4.3 million job postings and merging them to the Danish firm registry data. We document significant relationships between skill requirements and productivity, with LLM-based skill measures explaining three times more variation than keyword-based measures. The proposed framework demonstrates that the use of LLMs can sharply improve the extraction of economic variables from text data, while at the same time making such investigation much more widely accessible to economic researchers.

*Copenhagen Business School, Department of Economics; rz.eco@cbs.dk

†Copenhagen Business School, Department of Management, Society and Communication; dha.msc@cbs.dk

[1]This research was supported by a grant from The Rockwool Foundation (Grant 3034 – Digitalization: Jobs, firms and households), by the DeiC National Grant, and by the OpenAI Research Access Program. We thank Matthew Gentzkow, Ingar Haaland, and Anton Korinek for valuable comments and suggestions. Special thanks to Moira K. Daly, Cédric Schneider, Mathias Fjællegaard Jensen and Fane Groes for their invaluable contributions. We also extend our gratitude to Jenny Vismark and Martin Tacke Vismark for their excellent research assistance.

1 Introduction

Recent advances in large language models have opened new frontiers for extracting structured economic variables from unstructured text. Researchers have successfully deployed LLMs to decode policy documents and measure firm characteristics at unprecedented scales. Fang et al. [2025] extract policy objectives, tools, and implementation mechanisms from 3 million Chinese government documents, while Juhász et al. [2025] track industrial policy patterns across countries from 2009-2020. At the firm level, Li et al. [2025] quantify corporate culture by analyzing analyst reports and employee reviews. These applications align with Korinek [2023]'s framework for using generative AI in economic research, which identifies text analysis as key areas where LLMs can automate micro-tasks to achieve significant productivity gains in the economic research.

In labor economics, understanding how skill demands and workplace arrangements evolve has been central to explaining wage inequality, employment polarization, and gender gaps in the labor market (Acemoglu and Autor [2011]; Autor and Dorn [2013]; Goldin and Katz [2011];Hassan et al. [2025];Kalyani et al. [2025]). Job postings provide a direct window into these labor demand shifts, as Deming and Kahn [2018] demonstrated by documenting the rise of social skills. However, extracting structured measures from job posting text remains methodologically challenging. Early computational approaches made important advances: Atalay et al. [2020] applied the Continuous Bag of Words (CBOW) Model to track the long-run skill and task evolution in the U.S. Adams et al. [2019] used keywords and machine learning algorithms to measure workplace flexibility. Yet these methods are based on keywords and struggled with context-dependent meanings. The introduction of transformer architectures partially addressed these limitations. Hansen et al. [2023] achieved high accuracy in identifying remote work using BERT (Bidirectional Encoder Representations from Transformers), a natural language processing model from Google.

Building on the broader text-as-data literature in economics, we develop a comprehensive framework for measuring the full spectrum of skills and workplace flexibility attributes with large language models. This framework performs a multi-label classification task in the machine learning terminology [Madjarov et al., 2012, Bogatinovski et al., 2022]: each job posting typically exhibits 5-10 attributes simultaneously across our 25 categories. Our LLM-based approach addresses limitations of keyword methods identified by Gentzkow et al. [2019b], Ash and Hansen [2023], Hassan et al. [2025], particularly the challenge of capturing semantic meaning when terms have multiple interpretations depending on context. Our approach comprises four components: First, we develop a taxonomy of 25 classification categories grounded in economic literature (Deming and Kahn [2018], Adams et al. [2019], Nania et al. [2019]) and create detailed tagging instructions for each category. Second, we establish ground truth through manual an-

notation of 986 job postings by native Danish and English speakers. Third, we deploy this task to LLMs and evaluate both the GPT-4.1 family (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano) and open-source alternatives (Llama 3.3 70B, Qwen 2.5 7B). Fourth, we implement an iterative optimization process based on validation performance. When the accuracy or F1 score falls short, we employ three parallel strategies: refining prompts for greater specificity, tuning confidence thresholds to filter uncertain predictions, and fine-tuning models on the validation data.

In this paper, we find large language models outperform keyword methods for extracting economic variables from text. Analyzing 986 annotated Danish job postings across 25 categories, we find that GPT-4.1-mini slightly outperforms larger models like GPT-4.1 (accuracy: 0.824, Macro F1: 0.662) while open-source alternatives like Llama 3.3 70B achieve comparable performance (accuracy: 0.853, Macro F1: 0.652), enabling on-premise processing of confidential data without sacrificing classification quality. Besides, comparing different prompting strategies shows minimal variation in performance, with most approaches achieving similar accuracy around 0.853-0.860 and Macro F1 scores around 0.671-0.675, except for one strategy which underperforms at accuracy of 0.827 and Macro F1 of 0.639, suggesting that model architecture matters more than prompt engineering for this task. Beyond that, we have also implemented threshold tuning and fine-tuning methods. Threshold tuning leverages the inverse relationship between model confidence (log probabilities) and classification errors, establishing category-specific confidence thresholds that convert uncertain positive predictions to negatives. This method achieves up to 0.912 accuracy with macro F1 of 0.677, representing improvements over the baseline. Fine-tuning demonstrates even stronger performance gains through cross-validation experiments, with consistent improvements across all folds. Macro F1 scores ranging from 0.701 to 0.726 and accuracy consistently reaching 0.930 across all three folds, which representing improvement over keyword methods.

To demonstrate the economic relevance of these measurements, we also conduct a firm-level productivity analysis by matching our job posting data to Danish administrative records. Following Deming and Kahn [2018], we estimate the relationship between skill requirements and firm performance measured by revenue per employee. Our empirical application reveals that LLM-based skill measures capture economically meaningful variation that keyword methods fail to detect. The R-squared increases from 0.478 under keyword classification to 0.497 with LLM measures, demonstrating stronger explanatory power.

Section 2 describes our dataset and classification categories. Section 3 outlines our model selection process. Section 4 presents our human-in-the-loop framework, including validation annotation, baseline keyword approach, and LLM implementation. Section 5 reports optimization strategies and results across model selection, prompt engineering, threshold tuning, and fine-tuning. Section 6 presents the empirical application demonstrating economic significance of

LLM-based measurements. Section 7 concludes.

2 Data

2.1 Job Posting Data

The online job posting data analyzed in this paper is collected from the Danish labor market[2], posted from 2007 to 2023. The relevance and structure of this job posting data have been investigated by Jensen [2020], Bagger et al. [2022], Daly et al. [2022, 2025], who noted its similarities to the US job vacancy data from Burning Glass Technologies, extensively utilized in the research by Deming and Kahn [2018]. The data contains a firm identifier that can be matched to the Danish administrative data.

2.2 Skill and Flexibility Categories

We develop a taxonomy of 25 classification categories spanning digital skills, general skills, and workplace flexibility measures. We follow the digital classification developed by Nania et al. [2019] for Burning Glass Technologies for our digital skill categories. These categories were compiled from several works in the field. Deming and Kahn [2018] identified skill requirements across firms and labor markets through analysis of job postings, providing the foundation for our general skill categories. To avoid overlapping with digital skills, we removed the computer-related categories from Deming and Kahn [2018]'s categorization. Additionally, we follow Adams et al. [2019]'s framework for flexibility measures, who examined the relationship between different types of work flexibility and labor market outcomes, particularly in relation to gender wage gaps. The complete list of categories and their descriptions is provided in the Appendix: digital skills (Table 9), general skills (Table 10), and workplace flexibility measures (Table 11).

3 Large Language Model Selection

This paper evaluates the GPT-4.1 model family for this multi-label classification task. We selected the GPT-4.1 family comprising three variants: GPT-4.1 (flagship model), selected for state-of-the-art text understanding capabilities; GPT-4.1-mini (cost-effective with faster inference), chosen for 83% cost reduction and 50% lower latency while maintaining comparable performance; and GPT-4.1-nano (ultra-lightweight for high-volume processing), included for edge deployment scenarios as the fastest and cheapest option. All models support structured JSON output via response

[2]The Danish job posting data was constructed by the Danish consultancy firm Højbjerre Brauer Schultz Economics (HBS). Available at: https://hbseconomics.com/wp-content/uploads/2017/09/Eftersp%C3%B8rgslen-efterspr oglige-kompetencer.pdf

formatting, temperature control for deterministic outputs, and provide log probabilities for confidence assessment.

4 Method

4.1 Keyword Approach

The keyword approach, a foundational text-as-data method in economics, analyzes text by aggregating words or phrases while disregarding word order. We adopt this classical method as a benchmark. We implement a skill categorization framework using the comprehensive keyword list from Jensen [2020], Hassan et al. [2025], which builds on Deming and Kahn [2018]'s selective mapping strategy. Keywords are manually classified into skill categories or marked as noise, covering most occurrences. Synonyms from dictionary APIs extend this classification, with API definitions aiding uncategorized terms. Machine learning techniques (detailed in Appendix B.2 of Jensen [2020]) further refine the process. Skill extraction from job descriptions through keyword matching follows a three-stage process. First, the text undergoes preprocessing through lowercasing and punctuation removal. Next, the preprocessed text is matched against a predefined keyword list to count skill mentions. Finally, skill measurement occurs by aggregating category frequencies and compiling matched keywords for each job posting.

The keyword approach faces two fundamental limitations. First, it suffers from contextual blindness [Gentzkow et al., 2019a, Hassan et al., 2025]. Since the keyword approach lacks semantic modeling, it can lead to mis-classification when terms have multiple meanings. For instance, "leverage" could indicate management skills when used as "leverage your position" or financial expertise when referring to "leverage ratio analysis." Second, keyword methods can also generate false negatives. Job postings express requirements through varied vocabulary, which cannot fully be captured by predefined lists. These limitations motivate our LLM-based approach, which captures semantic meaning and identifies skills and flexibilities without being limited to a fix list of terms.

4.2 Large Language Model Approach

Figure 1 illustrates our systematic approach for extracting economic variables from unstructured text, which combines human expertise with LLM capabilities in an iterative refinement process. The framework begins with text cleaning and variable definition (Step 1), followed by the creation of detailed tagging instructions that human annotators use to generate ground truth data (Step 2). These human-annotated examples then guide the LLM in extracting variables from new text, with results parsed into structured format (Step 3). The critical innovation lies in Step 4's

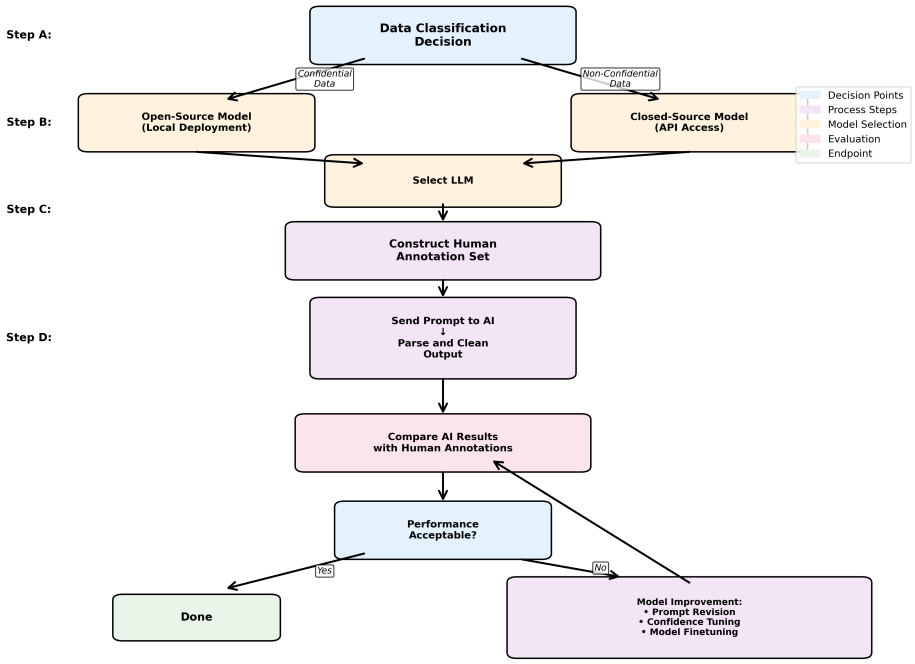

Figure 1: Human-in-the-Loop Framework for LLM-Based Multi-label Classification

evaluation loop: if the LLM's performance proves satisfactory against our validation set, the process completes; however, when improvements are needed, we employ three parallel optimization strategies. These include different prompt strategies (Section 5.1.3), tuning model confidence through false positive and negative analysis (Section 5.2.1), and fine-tuning the model with validation data (Section 5.2.2). This iterative approach ensures that our extraction method achieves acceptable accuracy while maintaining scalability. In the following sections, we detail each component of this framework, beginning with our evaluation metrics, then describing our validation annotation process, and finally presenting our LLM implementation and optimization strategies.

4.2.1 Evaluation Framework

We adopt the multi-label classification evaluation framework from Niraula et al. [2024], who provide systematic evaluations of generative LLMs for multi-label classification in technical domains. Following their approach, we employ macro-averaged metrics (F1-macro) and micro-averaged metrics (F1-micro) to assess model performance across our 25 classification categories. Macro-averaged metrics weight each category equally, providing insight into performance across imbalanced categories, while micro-averaged metrics weight each instance equally and are suitable for overall performance assessment. In multi-label classification where each label is predicted independently, accuracy is mathematically equivalent to the micro F1 score, as both metrics ag-

gregate true positives, false positives, and false negatives across all label-instance pairs before computing the final score. Therefore, we report only accuracy instead of micro F1 score to avoid redundancy. This evaluation framework enables direct comparison with existing multi-label classification benchmarks.

4.2.2 Validation Annotation Set

Following Ludwig et al. [2025]'s recommendation that LLM-based measurements require validation data, we construct a validation set by annotating 500 Danish and 500 English job postings, employing two native-speaker annotators (one per language). After removing job postings that were empty or failed quality checks, our final validation set contains 986 Danish job postings. Annotators followed detailed instructions defining skill measures and flexibility measures. The process consisted of two stages: first, annotators identified key terms, phrases, or sentences related to skill and flexibility measures. Then, they labeled the identified text by selecting the appropriate category names and tagging the corresponding content, similar to using a marker. To maintain quality, annotators relied solely on domain knowledge and external source checks (e.g., Google search) for technical term clarifications, excluding AI-generated suggestions. This approach ensured consistent categorization, yielding a reliable validation set for evaluating different LLM optimization strategies.

4.2.3 Model Deployment Considerations

LLMs are transformer-based neural networks that require substantial computational resources. Korinek [2024] provides detailed guidance on the practical aspects of LLM deployment, comparing API-based approaches with local deployment options and their respective trade-offs for economic research applications. For public data without privacy constraints, researchers can access commercial LLM APIs directly. However, when data contains personal information or confidential records, models must be deployed on local infrastructure or secure servers. This deployment constraint makes open-source models important for economic research. Our finding that the open-source model Llama-3.3-70B achieves performance parity with commercial APIs means researchers no longer face a trade-off between data security and classification quality.[3]

4.2.4 Model Configuration and Structured Output

We implement both configurations: OpenAI models are accessed through their API, while open-source models are deployed locally on NVIDIA H100 GPUs. We configure the models with three

[3]Commercial API deployment involves non-trivial costs that scale with dataset size. For our full empirical application, classifying 4.3 million job postings using GPT-4.1-mini through batch API totaled $3,447.

settings. First, we set temperature to 0, which means the model produces the least random outcomes when given the same input, minimizing randomness from predictions. Second, we enable the structured output function, which forces the model to return results in a format where each of the 25 categories receives a binary classification (1 for present, 0 for absent). Third, we enable log probabilities, which are numerical scores that indicate how confident the model is about each prediction. These confidence scores allow us to identify predictions where the model is uncertain, which we use later to improve classification accuracy.

4.2.5 Prompt Design

Our requirement for evidence extraction aligns with recent findings by Ma et al. [2025], who demonstrate that LLMs generate spiky probability distributions in multi-label classification, performing sequential single-label classification rather than holistic multi-label reasoning.

Our baseline prompt employs a three-component structure for skill and flexibility measurement extraction. The prompt begins with task instructions specifying the dual objective of identifying required job skills and workplace flexibility arrangements. The core input section presents the complete job description text. The output specification mandates a structured JSON response format containing 25 binary classification categories.

5 Large Language Model Optimization Strategies and Results

5.1 Model Selection and Prompt Engineering

5.1.1 Model Performance Analysis

Table 1 presents performance comparisons across model families for the 25-category classification task. The GPT family models demonstrate similar performance, with accuracy ranging from 0.824 to 0.853 and macro F1 from 0.627 to 0.673. Given that GPT-4.1-mini costs significantly less than GPT-4.1 while delivering comparable performance, we select GPT-4.1-mini as our default model for subsequent analyses.

Open-source models exhibit performance scaling with model size. Llama 3.3 70B achieves accuracy of 0.853 and macro F1 of 0.652, while Qwen 2.5 7B reaches accuracy of 0.798 and macro F1 of 0.092. This size-performance relationship indicates that larger open-source models are necessary in our experiment to match commercial API performance.

The keyword baseline serves as a natural benchmark against which to evaluate the incremental gains from large language models. Relative to this baseline (accuracy = 0.734, macro F1 = 0.425), large language models deliver clear performance improvements.

Table 1: Model Selection Performance Comparison

Model	Accuracy	Macro F1	Precision	Recall
GPT Family				
GPT-4.1-mini	0.853	**0.673**	0.649	0.810
GPT-4.1	0.824	0.662	0.606	0.855
GPT-4.1-nano	0.845	0.627	0.622	0.720
Open-Source				
Llama 3.3 70B	0.853	0.652	0.619	0.771
Qwen 2.5 7B	0.798	0.092	0.523	0.074
Keyword Baseline				
Keyword	0.734	0.425	0.461	0.586

Notes: Performance metrics computed on 986 manually annotated job postings across 25 classification categories. GPT-4.1 family models from OpenAI (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano) accessed via API with temperature=0 and structured JSON output. Open-source models Llama 3.3 70B (Meta AI) and Qwen 2.5 7B (Alibaba Cloud) deployed locally on NVIDIA H100 GPUs. All models evaluated using identical prompts and configuration settings. Keyword baseline implements the approach from Jensen [2020], Hassan et al. [2025].

5.1.2 Model Performance by Category

Tables 2 and 3 present the performance metrics. Performances are highly heterogeneous. For categories such as *characterSkills* (F1 = 0.916) and *socialSkills* (F1 = 0.906), the model attains near-human performance. By contrast, some other categories such as *flexibleTasks* and *GenAI* fall below standard thresholds.

Precision–recall trade-offs further reveal systematic classification biases. Categories with low accuracy, notably *flexibleTasks*, display high recall but low precision, consistent with systematic over-prediction. This motivates the threshold-tuning adjustment described in Section 5.2.1, which converts low-confidence positive predictions into negatives for these categories.

Table 2: Per-Category Performance Metrics: Skill Categories (Sorted by Accuracy)

Category	Accuracy	Precision	Recall	F1	Support
GenAI	0.995	1.000	0.167	0.286	6
AI	0.980	0.512	1.000	0.677	21
CRM	0.973	0.889	0.500	0.640	48
digitalMarketing	0.957	0.620	0.891	0.731	64
programmingAndSoftwareDevelopment	0.945	0.820	0.784	0.801	139
digitalDesign	0.926	0.670	0.573	0.618	103
machiningAndManufacturingTechnology	0.903	0.268	0.919	0.415	37
financialSkills	0.889	0.609	0.854	0.711	157
dataAnalysis	0.877	0.799	0.702	0.747	255
writingAndLanguage	0.863	0.802	0.983	0.883	518
computerSupportAndNetworking	0.863	0.701	0.359	0.475	170
characterSkills	0.848	0.849	0.994	0.916	822
managementSkills	0.841	0.736	0.871	0.798	356
socialSkills	0.834	0.830	0.997	0.906	795
customerService	0.789	0.687	0.865	0.766	393
productivity	0.716	0.476	0.807	0.599	259
cognitiveSkills	0.460	0.359	0.964	0.523	303

Table 3: Per-Category Performance Metrics: Flexibility Categories (Sorted by Accuracy)

Category	Accuracy	Precision	Recall	F1	Support
partTime	0.958	0.856	0.920	0.887	175
shiftWork	0.945	0.732	0.867	0.794	120
wfhOptions	0.939	0.821	1.000	0.902	275
temporary	0.926	0.663	0.925	0.773	134
flexibilityHoursEmployee	0.845	0.516	0.725	0.603	160
travelling	0.842	0.550	0.909	0.685	187
flexibilityHoursCompany	0.836	0.289	0.753	0.417	77
flexibleTasks	0.384	0.167	0.930	0.283	129

5.1.3 Prompt Engineering Strategy

Table 4: Prompt Engineering Strategies

Strategy	Description
Binary	Direct 0/1 prediction for each category.
Binary with Evidence (Baseline)	0/1 prediction with supporting text snippets, enabling intensity measures.
Occurrence count	Extract category mentions instead of binary decisions, approximating human annotation.
Two-shot	Separate skill and flexibility classifications into distinct API calls to reduce task complexity.

We evaluate four prompt strategies designed to test different aspects of LLM classification behavior. Binary classification provides the simplest baseline by requesting 1/0 decisions for each category. Binary classification with evidence extends this by requiring supporting text snippets, enabling construction of within-category intensity measures from extracted phrases. Occurrence-based scoring asks models to extract category mentions rather than make binary decisions, attempting to replicate human annotation behavior. Two-shot decomposition separates skills and flexibility classifications into distinct API calls, examining whether reducing task complexity through decomposition improves performance or whether the integrated nature of job postings makes joint classification more effective.

Table 5: Prompting Strategies Performance Comparison

Prompting Strategy	Accuracy	Macro F1	Precision	Recall
Binary Classification	0.860	0.675	0.638	0.799
Two-Shot Approach	0.859	0.675	0.632	0.802
Binary + Evidence	0.853	0.671	0.648	0.806
Occurrence Count	0.827	0.639	0.556	0.841

Notes: All strategies evaluated using GPT-4.1-mini on the same 986-posting validation set. *Binary Classification* requests direct 0/1 predictions for each category. *Two-Shot Approach* separates skills and flexibility classifications into two distinct API calls. *Binary + Evidence* adds requirement for supporting text snippets enabling within-category intensity measures (selected as baseline). *Occurrence Count* asks models to extract category mentions rather than binary decisions. Minimal performance variation across strategies indicates model architecture dominates prompt design for this classification task.

From Table 5, the minimal performance differences across strategies indicate that prompt complexity has limited impact on classification outcomes. Binary classification achieves accuracy

of 0.860 and macro F1 of 0.675, while the two-shot approach reaches nearly identical scores at 0.859 accuracy and 0.675 macro F1 despite doubling API calls. Binary classification with evidence shows slightly lower performance at 0.853 accuracy and 0.671 macro F1. Occurrence counting performs worst with 0.827 accuracy and 0.639 macro F1, indicating that occurrence-based approaches mis-align with the binary classification task. This paper adopts the binary classification with evidence strategy as the default approach to enable within-category intensity measurement through extracted text phrases.

5.2 Performance Optimization

5.2.1 Log Probability Threshold Tuning

LLMs provide log probabilities representing token generation likelihood. Ivanova et al. [2024], Kauf et al. [2024] demonstrates that log probability scores measure model confidence and align predictions with human judgments. Values closer to zero indicate higher confidence. We use these scores to improve accuracy by filtering uncertain predictions through category-specific confidence thresholds that convert low-confidence positive predictions to negative ones, reducing false positives. Our use of log probabilities to filter uncertain predictions also builds on Ma et al. [2025]'s analysis of LLM multi-label mechanisms, which shows that models exhibit spiky distributions with each step strongly favoring single labels. Their finding that relative label rankings poorly predict subsequent predictions motivates our category-specific threshold approach.

Figure 2 plots the relationship between category-level log probability scores and human-AI disagreement rates. Five categories exhibit high disagreement rates exceeding 20 percent: *cognitiveSkills*, *flexibleTasks*, *productivity*, *customerService*, and *managementSkills*. These categories show strong negative correlation between log probability values and disagreement rates, indicating that model uncertainty aligns with classification errors. Categories clustering near zero log probability with disagreement rates below 10 percent demonstrate reliable baseline performance. These include *wfhOptions*, *CRM*, and most digital skill categories.

Based on the finding, we develop a novel confidence-based threshold tuning method that exploits the relationship between model uncertainty and classification errors to improve measurement accuracy without additional training. Rather than accepting all positive predictions at face value, our method establishes category-specific confidence thresholds using log probabilities, converting uncertain positive predictions to negative classifications. This asymmetric treatment directly addresses the false positive problem that dominates classification errors. The economic significance extends beyond performance metrics: threshold tuning enables researchers to calibrate measurement instruments for specific research objectives—prioritizing precision or recall.

Table 6 reports three-fold cross-validation results showing that threshold tuning produces ac-

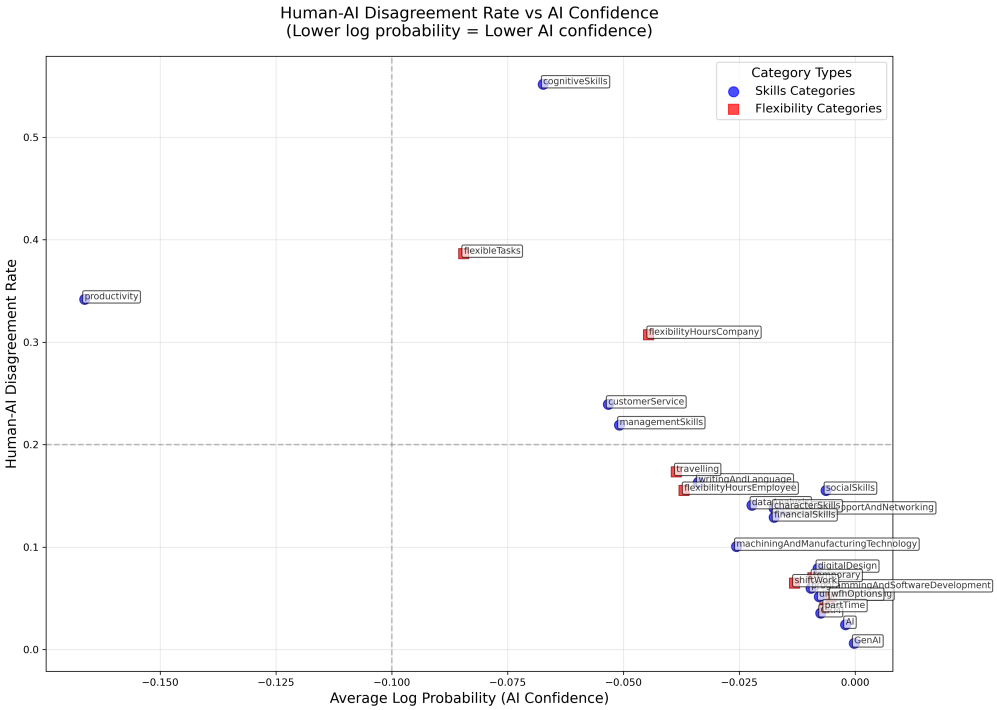

Figure 2: Disagreement vs. Log Probability by category

Table 6: Three-Fold Threshold Tuning Performance

Model	Accuracy	Macro F1	Precision	Recall
GPT-4.1-mini (Baseline)	0.853	0.673	0.649	0.810
Threshold Tuning Fold 3	0.912	0.677	0.700	0.715
Threshold Tuning Fold 2	0.903	0.639	0.649	0.690
Threshold Tuning Fold 1	0.903	0.632	0.633	0.702

Notes: Threshold tuning exploits the inverse relationship between model log probabilities and classification errors. Category-specific confidence thresholds convert low-confidence positive predictions to negatives, reducing false positives without additional training. Each fold uses approximately 657 postings to establish thresholds and 329 for testing.

curacy gains (from 0.853 to 0.912 in the best fold) while maintaining comparable macro F1 scores. Macro F1 ranges from 0.632 to 0.677 across folds compared to the baseline of 0.673, reflecting the precision-recall trade-off inherent in threshold optimization. The method proves most effective for problematic categories like *FlexibleTasks* and *CognitiveSkills*, where disagreement rates fall by over 20 percentage points. This result establishes threshold tuning as a practical optimization technique for LLM-based economic measurement, especially when one of the false positive or false negative dominates the error.

5.2.2 Model Fine-tuning

Fine-tuning LLMs adapts pre-trained language models to specific tasks by updating their internal neural network weights through additional training data Howard and Ruder [2018]. Unlike the base model trained on general text, fine-tuning adjusts the model's parameters to optimize performance for the target classification task. This process preserves the model's general language understanding while specializing its outputs for job posting classification. Also, implementation complexity varies significantly between deployment methods. Commercial API fine-tuning through OpenAI requires only data formatting and API calls, completing training in approximately 2 hours per model. Open-source fine-tuning demands substantially more resources: configuring training infrastructure, managing GPU memory allocation, implementing gradient accumulation strategies, and monitoring convergence. These technical requirements and computational costs make API-based fine-tuning more accessible for most research applications in the field of economics, though open-source alternatives remain necessary for sensitive and confidential data.

The training data structure follows a format that mirrors human annotation practices. Each training example consists of an input-output pair where the input contains the complete job posting text followed by a template specification for each skill category, formatted as placeholder brackets indicating the expected output structure. The corresponding output contains the actual human annotations, with binary classifications (0 or 1) paired with evidence text extracted directly by human annotators when a skill is present. For instance, when annotators identify CRM skills in a posting, the output records both the positive classification and the specific text fragment supporting this judgment, such as "CRM ERP" or "Salesforce experience required."

We finetune the model gpt-4.1-2025-04-14 through OpenAI's API. The training process employs three-fold cross-validation, with each fold using approximately 657 postings for training and 329 for testing. Table 7 demonstrates that fine-tuning achieves the highest performance among all optimization strategies. The fine-tuned models reach average accuracy of 0.930 and macro F1 of 0.712 across folds. The minimal cross-fold variation confirms that performance gains are persistent. While fine-tuning requires upfront computational investment, it provides optimal

classification performance.

Table 7: Three-Fold Cross-Validation Finetuning Performance

Model	Accuracy	Macro F1	Precision	Recall
GPT-4.1-mini (Baseline)	0.853	0.673	0.649	0.810
Cross-Validation Fold 3	0.929	0.726	0.773	0.704
Cross-Validation Fold 2	0.930	0.708	0.744	0.682
Cross-Validation Fold 1	0.930	0.701	0.728	0.693
CV Average	0.930	0.712	0.748	0.693

Notes: Fine-tuning results for GPT-4.1 conducted via OpenAI API. Each fold uses approximately 657 postings for training and 329 for testing. Training data pairs complete job posting text with human annotations including binary classifications and evidence text extracted by annotators.

6 Empirical Application: Economic Significance of LLM-Based Measurements

To demonstrate the economic relevance of our LLM-based skill measurements and compare their performance against keyword approaches, we conduct a firm-level productivity analysis. This application serves two objectives: first, establishing that LLM-derived measures capture economically meaningful variation in labor demand; second, directly comparing the explanatory power of LLM versus keyword classifications.

6.1 Data and Specification

We match our job posting data to Danish administrative firm records using firm identifiers, obtaining revenue and employment information for vacancy-posting firms. Our primary performance measure is log revenue per employee, capturing firm-level labor productivity.

Following Deming and Kahn [2018], we estimate:

$$\text{Log Revenue per Employee}_f = \alpha_0 + \overline{\text{Skill}}_f \beta' + \overline{I}^o_f + \overline{X}_f \gamma' + \theta_n + \epsilon_f \tag{1}$$

where $\overline{\text{Skill}}_f$ represents firm-level average shares of postings requiring each skill category (computed separately using LLM and keyword methods), $\overline{I}^o_f$ denotes occupation shares, $\overline{X}_f$ captures average education levels, and θ_n represents industry fixed effects. Regressions are weighted by each firm's vacancy count.

6.2 Results

Table 8 presents results across three specifications. Column (1) provides a baseline specification without skill measures, Column (2) employs keyword-based skill classifications, and Column (3) uses LLM-derived skill measures. All specifications include firm-level controls for age and size, occupation shares, average education levels, and industry fixed effects.

The LLM-based skill measures reveal economically meaningful relationships with firm productivity that keyword methods largely fail to capture. Column 3 of Table 8 demonstrates that social skills exhibit the strongest positive association with revenue per employee, with a coefficient of 0.536 suggesting that 10 percentage point increase in social skill requirements corresponds to approximately 5 percentage higher labor productivity. Cognitive skills similarly show positive coefficient of 0.248. The keyword approach yields a small negative coefficient of -0.013 for social skills and fails to identify significant effects for cognitive skills. The attenuated point estimates near zero in the keyword specification likely reflect classical measurement error in skill classification. In other words, when job advertisements use varied or indirect language to describe skill requirements, keyword matching produces noisy measures that bias coefficients toward zero. The LLM approach appears to correct these measurement issues.

The superiority of LLM methods extends beyond individual coefficients to overall model performance. The R-squared increases from 0.478 under keyword classification to 0.497 with LLM measures. This improvement in explanatory power potentially stems from the LLM's ability to detect nuanced skill requirements that simple keyword matching overlooks. The consistent pattern of larger, more significant coefficients indicates that LLMs capture meaningful variation in labor demand that traditional text analysis methods miss, validating the use of these advanced techniques for understanding firm-level skill requirements and their economic consequences.

7 Conclusion

This paper establishes large language models as measurement instruments for extracting economic variables from unstructured text. Through evaluation on 986 annotated Danish job postings across 25 skill and flexibility categories, we demonstrate that LLMs transform labor market text data into structured measures with accuracy and economic relevance.

Our findings carry four methodological implications for researchers. First, base LLMs outperform keyword methods without requiring task-specific training, achieving 85.3% accuracy versus 73.4% for keywords and macro F1 scores of 0.673 versus 0.425. This performance gap reflects LLMs' ability to capture semantic context rather than relying on string matching. Second, prompt engineering delivers gains of less than three percentage points in accuracy across strategies. Model architecture, not prompt design, determines classification quality for tasks with

Table 8: Regression Results: Log Revenue per Person

Variable	(1) No Skill	(2) Keyword	(3) LLM
Firm Age	0.0041***	0.0044***	0.0038***
	(0.0005)	(0.0005)	(0.0005)
Firm Size	-0.0035	-0.0071	-0.0005
	(0.0041)	(0.0043)	(0.0044)
Cognitive Skill		-0.0038	0.2475***
		(0.0049)	(0.0615)
Social Skill		-0.0133***	0.5358***
		(0.0032)	(0.0894)
Customer Service		-0.0010	-0.3567***
		(0.0017)	(0.0557)
Character Skill		0.0021	-0.7256***
		(0.0017)	(0.0907)
Financial Skill		0.0003	0.0356
		(0.0042)	(0.0801)
Management Skill		0.0208***	0.1322
		(0.0032)	(0.0701)
Writing and Language		0.0049	-0.1520**
		(0.0037)	(0.0472)
CRM		-0.1142***	-1.0355***
		(0.0319)	(0.1915)
Computer Support		0.0197	0.5050***
		(0.0128)	(0.1155)
Data Analysis		0.0162	0.2442*
		(0.0497)	(0.1206)
Digital Design		-0.1573*	-1.2872***
		(0.0674)	(0.1694)
Digital Marketing		-0.0275*	0.2257*
		(0.0111)	(0.1065)
Machining Manufacturing		-0.2323***	-0.3185***
		(0.0659)	(0.0762)
Productivity		-0.0118	0.2991***
		(0.0343)	(0.0743)
Programming Software		-0.0151	0.1986
		(0.0127)	(0.1264)
Observations	6,566	6,566	6,566
R-squared	0.4700	0.4780	0.4970
F-statistic	99.3280	81.4850	87.7240

Standard errors in parentheses. Clustered standard errors applied.
All regressions control for firm-level characteristics, including occupation shares (I_{of}) and average education levels (X_f). Industry fixed effects (θ_n) are included. Regressions are weighted by each firm's vacancy count.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

specification. Researchers can prioritize task specification over prompt optimization.

Third, we document an inverse relationship between model confidence and classification errors. This relationship enables a threshold-tuning approach that achieves 91.2% accuracy by converting predictions with confidence to negatives based on log probability scores. Threshold tuning requires no training and allows researchers to calibrate measurement precision based on priorities. Categories with baseline error rates, such as flexible tasks and cognitive skills, show disagreement reductions exceeding 20 percentage points through threshold adjustment. Fourth, fine-tuning reaches performance at 93% accuracy with 0.712 macro F1 across cross-validation folds, though at cost. The choice between threshold tuning and fine-tuning involves trade-offs between implementation complexity and accuracy.

A contribution lies in demonstrating performance parity between APIs and open-source alternatives. Llama-3.3-70B matches GPT-4.1-mini's classification quality while enabling on-premise deployment, addressing data confidentiality constraints that have limited LLM adoption in economics. This finding expands researcher access to text processing for records, firm-level data, and information that cannot be transmitted to servers.

The economic significance of LLM-based measurements extends beyond classification metrics. Our firm-level productivity analysis reveals that LLM-derived skill measures explain variation in revenue per employee compared to keyword approaches, with R-squared increasing from 0.478 to 0.497. LLM methods detect relationships—such as the 54 percent productivity premium associated with social skill requirements—that keyword methods fail to identify. These results validate LLM measurements as capturing economic variation rather than achieving classification accuracy.

The framework developed here generalizes beyond job postings. The human-in-the-loop approach combining validation annotation, baseline evaluation, and optimization applies to measurement challenges: extracting innovation indicators from patents, identifying policy instruments in government documents, measuring sentiment in disclosures, or coding survey responses. As LLM capabilities expand and costs decline, these methods will become tools for converting unstructured text into structured economic data. This democratization of text analysis creates opportunities for addressing research questions constrained by measurement limitations, in settings where coding schemes prove insufficient or where data volume renders manual annotation infeasible.

Research should examine LLM robustness across languages, time periods, and document types, and explore allocation of resources between human annotation and model training. The costs and capabilities of language models suggest that text-based economic measurement will continue to evolve, opening frontiers where language provides the window into economic phenomena.

A Step A: Data Classification Categories

Category	Description
Artificial Intelligence (AI)	Skills related to AI technologies and machine intelligence.
Generative AI (GenAI)	Skills related to generating content using AI models.
CRM	Skills involving customer relationship management.
Computer Support and Networking	Skills in setting up, managing, and supporting computer systems and networks.
Data Analysis	Skills in analyzing, visualizing, and interpreting data.
Digital Design	Skills in creating digital content, including UI/UX and multimedia design.
Digital Marketing	Skills related to online marketing and analytics.
Machining and Manufacturing Technology	Skills related to manufacturing processes and technical engineering.
Productivity	Skills in tools that improve workflow and collaboration.
Programming and Software Development	Skills in software programming, development, and related tools.

Table 9: Digital Skills Categories

Category	Description
Character Traits	Personal qualities or work habits that contribute to job performance.
Cognitive Skills	Intellectual and analytical skills for problem-solving.
Customer Service	Skills related to interacting with customers or clients.
Financial Skills	Skills in financial tasks like budgeting and accounting.
Management Skills	Skills required for team leadership, supervision, and resource management.
Social Skills	Interpersonal skills for effective communication and collaboration.
Writing and Language	Skills related to writing proficiency and language abilities.

Table 10: General Skills Categories

Category	Description
Working from Home (wfhOptions)	Option to work outside a traditional office setting.
Company-Driven Flexibility Hours	Requires scheduling flexibility based on company needs.
Employee-Driven Flexibility Hours	Allows employees to set their own hours.
Flexible Tasks	Expectation for adaptability in tasks and responsibilities.
Temporary	Limited-duration or project-based positions.
Part Time	Position with reduced hours compared to full-time.
Shift Work	Work requiring non-standard or rotating shifts.
Travelling	Role requires regular travel or working across multiple locations.

Table 11: Workplace Flexibility Categories

B Step B: Model Selection Specifications

B.1 Open-Source vs Closed-Source Model Criteria

- **Confidential Data**: Use open-source models with local deployment
- **Non-Confidential Data**: Use closed-source models with API access

C Step C: LLM Selection and Prompt Specifications

C.1 Selected LLM Models

- GPT-4.1-nano
- GPT-4.1-mini
- GPT-4.1

C.2 Complete Base Prompt Text

```
Analyze this job posting to identify required skills and
workplace flexibility options.

Job Description:
{job_description}

Return ONLY valid JSON in this exact format:
{
    "skills": {
        "AI": [1/0, "evidence"],
        "GenAI": [1/0, "evidence"],
        "CRM": [1/0, "evidence"],
        "computerSupportAndNetworking": [1/0, "evidence"],
        "dataAnalysis": [1/0, "evidence"],
        "digitalDesign": [1/0, "evidence"],
        "digitalMarketing": [1/0, "evidence"],
        "machiningAndManufacturingTechnology": [1/0, "evidence"],
        "productivity": [1/0, "evidence"],
        "programmingAndSoftwareDevelopment": [1/0, "evidence"],
        "characterSkills": [1/0, "evidence"],
        "cognitiveSkills": [1/0, "evidence"],
```

```
        "customerService": [1/0, "evidence"],
        "financialSkills": [1/0, "evidence"],
        "managementSkills": [1/0, "evidence"],
        "socialSkills": [1/0, "evidence"],
        "writingAndLanguage": [1/0, "evidence"]
    },
    "flexibility": {
        "wfhOptions": [1/0, "evidence"],
        "flexibilityHoursCompany": [1/0, "evidence"],
        "flexibilityHoursEmployee": [1/0, "evidence"],
        "flexibleTasks": [1/0, "evidence"],
        "temporary": [1/0, "evidence"],
        "partTime": [1/0, "evidence"],
        "shiftWork": [1/0, "evidence"],
        "travelling": [1/0, "evidence"]
    },
    "valid": [1/0, "reason if 0"]
}

Instructions:
- Return 1 if present, 0 if not
- Include key evidence phrases when found
- Set "valid" to 0 if not a job posting
- OUTPUT ONLY JSON, NO EXPLANATIONS OR REASONING
```

C.3 Human Annotation Platform

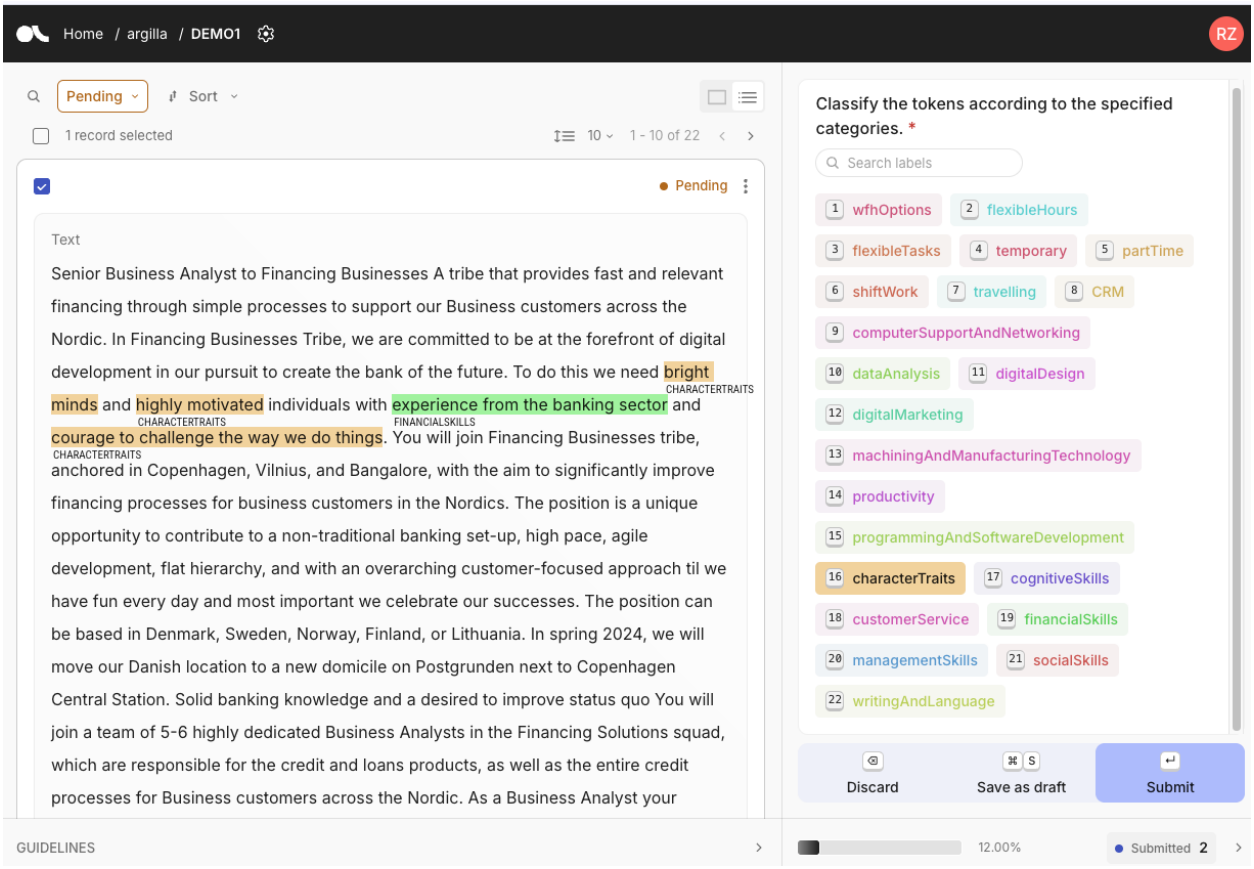

Figure 3: Human Annotation Interface Demo

D Step D: AI Processing and Model Improvement

D.1 Model Improvement Techniques

- **Prompt Revision**: Iterative refinement of prompt structure and instructions
- **Confidence Tuning**: Threshold adjustment using log probability analysis
- **Model Finetuning**: Training specialized models on domain-specific data

D.2 Category Definitions and Keyword Examples

D.2.1 Skills Categories

Table 12: Skills Categories: Digital Skills with Bilingual Keyword Examples

Category	English Keywords	Danish Keywords
AI	machine learning, neural networks, tensorflow, pytorch, artificial intelligence, deep learning, ml, ai	maskinlæring, neurale netværk, kunstig intelligens
GenAI	chatgpt, generative ai, prompt engineering, llm, large language model, gpt, openai	generativ ai, kunstig intelligens
CRM	salesforce, hubspot, crm software, customer relationship management	kundeservice system, crm
Computer Support	it support, cybersecurity, network administration, technical support, helpdesk, networking	it support, netværk, cybersikkerhed, teknisk support
Data Analysis	python, sql, tableau, data analysis, analytics, statistics, r programming, excel	dataanalyse, statistik, excel, python, sql
Digital Design	ui/ux, adobe suite, figma, photoshop, illustrator, graphic design, web design	grafisk design, webdesign, adobe, photoshop
Digital Marketing	seo, google analytics, social media marketing, digital marketing, sem, ppc	digital marketing, sociale medier, seo, online marketing
Manufacturing Tech	cnc, cad/cam, plcs, automation, manufacturing, machining	cnc, automatisering, produktion, maskiner
Productivity	microsoft office, project management tools, excel, powerpoint, word, outlook	microsoft office, excel, word, powerpoint, projektledelse
Programming	programming, javascript, react, api development, software development, coding	programmering, softwareudvikling, kodning, javascript

Table 13: Skills Categories: General Skills with Bilingual Keyword Examples

Category	English Keywords	Danish Keywords
Character Skills	organized, reliable, adaptable, detail-oriented, self-motivated, proactive	organiseret, pålidelig, fleksibel, selvmotiveret, struktureret
Cognitive Skills	critical thinking, problem-solving, analytical, decision making, troubleshooting	problemløsning, analytisk, kritisk tænkning, fejlfinding
Customer Service	customer service, client relations, customer support, client management	kundeservice, kundekontakt, kundesupport, kundebehandling
Financial Skills	budgeting, financial analysis, accounting, finance, bookkeeping	økonomi, budgettering, regnskab, bogføring, finansiel analyse
Management	leadership, team management, supervision, project management	ledelse, teamledelse, projektledelse, lederskab, supervision
Social Skills	communication, teamwork, collaboration, interpersonal, team player	kommunikation, teamwork, samarbejde, sociale færdigheder
Writing/Language	writing, editing, language proficiency, content creation, technical writing	skrivning, sprogfærdigheder, kommunikation, tekstforfattning

D.2.2 Workplace Flexibility Categories

Table 14: Workplace Flexibility Categories with Bilingual Keyword Examples

Category	English Keywords	Danish Keywords
Work from Home	remote work, hybrid, telecommuting, work from home, wfh, remote	hjemmearbejde, remote arbejde, fleksibel arbejdsplads
Company Hours	on-call, rotating shifts, variable hours, flexible schedule, shift work	skiftarbejde, fleksible timer, varierende arbejdstid
Employee Hours	flexible hours, self-managed schedule, choose your hours, flexible timing	fleksible arbejdstider, selvbestemt arbejdstid, fri tilrettelæggelse
Flexible Tasks	varied responsibilities, cross-functional, diverse tasks, multiple duties	varierende opgaver, forskellige ansvarsområder, alsidige opgaver
Temporary Work	contract, seasonal, project-based, temporary, temp, short-term	kontrakt, midlertidig, projekt, sæsonarbejde, vikar
Part Time	part-time, reduced hours, part time, flexible hours	deltid, deltidsstilling, reducerede timer
Shift Work	shift work, night shift, evening shift, rotating shifts	skiftarbejde, natarbejde, aftenarbejde, weekendarbejde
Travelling	business travel, field work, travel required, on-site work	rejsearbejde, feltarbejde, kørsel, udstationering

References

D Acemoglu and DH Autor. Chapter 12-skills, tasks and technologies: Implications for employment and earnings (d. card & o. ashenfelter, eds.). *Elsevier. https://doi. org/10.1016/S0169-7218 (11)*, pages 02410–5, 2011.

Abi Adams, Mathias Fjællegaard Jensen, and Barbara Petrongolo. The contribution of employee-led and employer-led work flexibility to the motherhood wage gap. 2019.

Elliott Ash and Stephen Hansen. Text Algorithms in Economics. *Annual Review of Economics*, 15(1):659–688, September 2023. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-082222-074352. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-082222-074352. Publisher: Annual Reviews.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. The evolution of work in the united states. *American Economic Journal: Applied Economics*, 12(2):1–34, 2020.

David H Autor and David Dorn. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*, 103(5):1553–1597, August 2013. ISSN 0002-8282. doi: 10.1257/aer.103.5.1553. URL https://pubs.aeaweb.org/doi/10.1257/aer.103.5.1553. Publisher: American Economic Association.

Jesper Bagger, François Fontaine, Manolis Galenianos, and Ija Trapeznikova. Vacancies, employment outcomes and firm growth: Evidence from denmark. *Labour Economics*, 79:102138, 2022. doi: 10.1016/j.labeco.2021.102138.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203: 117215, 2022.

Moira K. Daly, Mathias Fjællegaard Jensen, and Daniel le Maire. University admission and the similarity of fields of study: Effects on earnings and skill usage. *Labour Economics*, 75:102141, April 2022. doi: 10.1016/j.labeco.2022.102141.

Moira K. Daly, Fane Naja Groes, and Mathias Fjællegaard Jensen. Skill demand versus skill use: Comparing job posts with individual skill use on the job. *Labour Economics*, 92:102640, January 2025. doi: 10.1016/j.labeco.2024.102640.

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

Hanming Fang, Ming Li, and Guangli Lu. Decoding China's Industrial Policies. *SSRN Electronic Journal*, 2025. ISSN 1556-5068. doi: 10.2139/ssrn.5078043. URL https://www.ssrn.com/abstract=5078043. Publisher: Elsevier BV.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019a.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57(3):535–574, September 2019b. ISSN 0022-0515. doi: 10.1257/jel.20181020. URL https://pubs.aeaweb.org/doi/10.1257/jel.20181020. Publisher: American Economic Association.

Claudia Goldin and Lawrence F. Katz. The Cost of Workplace Flexibility for High-Powered Professionals. *The ANNALS of the American Academy of Political and Social Science*, 638(1): 45–67, November 2011. ISSN 0002-7162, 1552-3349. doi: 10.1177/0002716211414398. URL https://journals.sagepub.com/doi/10.1177/0002716211414398.

Stephen Hansen, Peter John Lambert, Nicholas Bloom, Steven J Davis, Raffaella Sadun, and Bledi Taska. Remote work across jobs, companies, and space. Technical report, National Bureau of Economic Research, 2023.

Tarek A. Hassan, Stephan Hollander, Aakash Kalyani, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun. Text as data in economic analysis. *Journal of Economic Perspectives*, 39 (3):193–220, 2025.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Evelina Fedorenko, and Jacob Andreas. Log probability scores provide a closer match to human plausibility judgments than prompt-based evaluations. In *Proceedings of the South NLP Symposium 2024*, Atlanta, GA, April 2024. Emory University. URL https://southnlp.github.io/southnlp2024/papers/southnlp2024-poster-47.pdf.

Mathias Fjællegaard Jensen. Gender differences in returns to skills. In *3rd IDSC of IZA/University of Luxembourg Workshop*, 2020.

Réka Juhász, Nathan Lane, Emily Oehlsen, and Verónica Pérez. The Who, What, When, and How of Industrial Policy: A Text-Based Approach. 2025.

Aakash Kalyani, Nicholas Bloom, Marcela Carvalho, Tarek Hassan, Josh Lerner, and Ahmed Tahoun. The diffusion of new technologies. *The Quarterly Journal of Economics*, 140(2): 1299–1365, 2025.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 264–286. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.18. URL https://aclanthology.org/2024.blackboxnlp-1.18/.

Anton Korinek. Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, 61(4):1281–1317, December 2023. ISSN 0022-0515. doi: 10.1257/jel.20231736. URL https://pubs.aeaweb.org/doi/10.1257/jel.20231736. Publisher: American Economic Association.

Anton Korinek. Generative ai for economic research: Llms learn to collaborate and reason. Technical report, National Bureau of Economic Research, 2024.

Kai Li, Feng Mai, Rui Shen, Chelsea Yang, and Tengfei Zhang. Dissecting Corporate Culture Using Generative AI. 2025.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework, 2025.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently. *arXiv preprint arXiv:2505.17510*, 2025.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.

J. Nania, H. Bonella, D. Restuccia, and B. Taska. No longer optional: Employer demand for digital skills, 2019.

Nobal Niraula, Samet Ayhan, Balaguruna Chidambaram, and Daniel Whyatt. Multi-label classification with generative large language models. In *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, pages 1–7, Sep. 2024. doi: 10.1109/DASC62030.2024.10748883.