Skill Substitutability is Contextual: Extracting Dynamic Skill Substitutes from Job Postings

Anonymous ACL submission

Abstract

Finding suitable substitutes in a specific context is a core challenge for recommendation and information retrieval systems. In the labor market, identifying appropriate skill substitutes can alleviate skill mismatch—the discrepancy between employer demand and workforce supply. Yet existing context-agnostic methods capture relatedness rather than functional interchangeability. We propose a framework that fine-tunes a Masked Language Model (BERT) on job postings to extract contextual substitutes from the predicted probability distribution over a fixed skill vocabulary. Our model substantially outperforms word2vec, conditional probability, and GPT-5.1 baselines. Importantly, we find that our model matches HR professionals' substitutability judgments, tested on job postings where interchangeable skills are explicitly stated. While substitutability is highly dynamic across time and occupations simultaneously, our investigation using SAE reveals that deeper layers increasingly organize skill representations along context-specific dimensions, offering a representational account of how our model successfully captures the dynamics. Our simulation further confirms that using context-aware, trending substitutes can increase job posting visibility by up to three orders of magnitude. We expect that our approach offers a scalable, context-aware methodology for navigating the rapidly evolving labor market.

1 Introduction

Finding the most suitable substitute in a specific context, for a given item, has been a significant challenge in computer science, contributing to the advancement in web search, information retrieval, and recommendation systems (Finkelstein et al., 2001; Kraft et al., 2005, 2006; Rakesh et al., 2019; Rashed et al., 2022). When a user's initial query fails or an item is unavailable, the ability to offer a contextually relevant alternative is crucial for a successful user experience. In this regards, previous studies on finding the best substitute have evolved towards the following direction: from global, static notions of similarity to dynamic, context-dependent substitutability.

Extracting the best substitute is particularly essential in the labor market, where identifying the right substitute for a given skill has been a long-standing issue for economists, policymakers, and businesses. While economists have struggled with how to measure and value the skills of the workforce (Wasmer, 2001; Ingram and Neumann, 2006; Quintini, 2014; Borghans et al., 2014; Deming and Kahn, 2018; Guvenen et al., 2020), a persistent challenge from this inquiry is "skill mismatch"—a discrepancy between the skills employers demand and those that the workforce possesses (Handel, 2003; Weaver and Osterman, 2017; McGuinness et al., 2018).

Due to its direct and detrimental effects on both micro- and macroeconomic outcomes, understanding and resolving skill mismatch is critical in the national economy. For individuals, a mismatch is often associated with wage penalties and lower job satisfaction (Allen and Van der Velden, 2001; Green and Zhu, 2010; Liu et al., 2016; Guvenen et al., 2020). For firms, it can lead to unfilled vacancies, higher training costs, and reduced productivity (Fanti et al., 2021; Rathelot and Thijs Chan, 2023; Florez and Gomez, 2024; McGuinness et al., 2018). At the national level, a high degree of skill mismatch can result in structural unemployment and act as a significant drag on economic growth and innovation (McGuinness et al., 2018).

A dynamic and contextual understanding of skill substitutability can directly address this problem. For instance, a job posting for a data analyst might explicitly require proficiency in the "R" programming language. A job seeker who is an expert in "Python" with relevant data science libraries might

be overlooked by an automated screening system, even though their skills are highly substitutable for the required tasks. A system capable of recognizing this contextual interchangeability would create a more efficient matching process, benefiting both employers and job seekers (Kim et al., 2024).

Recent computational approaches have introduced new ways to characterize the structure of the labor market. Prior work has applied network analysis to model relationships among skills, occupations, and industries (Anderson, 2017; Alabdulkareem et al., 2018; Park et al., 2019). In parallel, a growing body of research has leveraged natural language processing (NLP) methods to track changes in skill demand over time, particularly in response to technological change (Hershbein and Kahn, 2018; Deming and Noray, 2020; Alekseeva et al., 2021; Evans et al., 2024; Ham et al., 2025). More recently, word embedding models and large language models such as BERT have been used to learn distributed representations not only of skills, but also of other labor-market entities, including occupations, industries, and firms (Tong et al., 2021; Kim et al., 2024; Van-Duyet et al., 2017).

However, while these studies have excelled at mapping the static and global relationships between skills, they have not been designed to capture the dynamic, context-dependent nature of substitutability. Their reliance on aggregated co-occurrence data shows that they primarily model skill relatedness or similarity, which is not the same as interchangeability in a specific job function. The novelty of our work lies in shifting the focus from "what skills are related?" to "what skill is the best substitute for this specific occupation?"

To answer this question, we introduce a framework that extracts contextual skill substitutes by fine-tuning a Masked Language Model (BERT) on millions of job postings (2010–2025). By predicting masked skills from the output probability distribution, our approach captures fine-grained interchangeability and substantially outperforms context-agnostic baselines, including *word2vec* and *GPT-5.1*. Most notably, our model's predicted substitutes closely match the skill alternatives that HR professionals explicitly mark as interchangeable in real job postings, reaching near-expert accuracy despite being far more lightweight than a general-purpose LLM. We further find that substitutability is highly dynamic across occupations and over time, and a Sparse Autoencoder (SAE) analysis (Bricken et al., 2023; Wu et al., 2024) shows that deeper layers increasingly separate context-specific signals, offering a representational account of how the model captures this variation. Finally, simulations confirm that context-aware, trending substitutes can increase job visibility by up to three orders of magnitude, providing a scalable tool for labor market navigation.

2 Methods

Building on a masked language modeling framework originally developed for detecting implicit dehumanizing rhetoric (Card et al., 2022), we repurpose this objective to recover context-appropriate substitutes for workplace skills. As illustrated in Figure 1, our approach allows us to measure semantic proximity and reveal substitutability patterns that static co-occurrence statistics fail to capture. We focus our analysis on IT and software development postings, as these rapidly evolving skills provide an ideal testbed for examining contextual substitutability.

Given a job posting description $X = (x_1, \ldots, x_{[MASK]}, \ldots, x_N)$ in which a target skill s^* is replaced with [MASK], the model predicts s^* among fixed vocabulary of candidate skills $\mathcal{V}$. Formally, the model outputs a probability distribution $P(s \mid X)$ over $s \in \mathcal{V}$. We consider alternative skills with high probability in $P(s \mid X)$ as plausible contextual substitutes for the target skill s^*.

2.1 Job Postings Dataset

We use longitudinal job posting data from *Lightcast*, a large-scale labor-market corpus covering more than 3.7 billion U.S. postings from 2010 to 2025 across diverse industries. The dataset includes a standardized skill taxonomy for consistent skill identification and 139 structured variables grouped into metadata, job classification, and skills & certifications categories (Appendix Table A.4), enabling analysis of long-term skill dynamics, technology diffusion, and temporal substitution patterns (Deming and Kahn, 2018).

2.2 Preprocessing

Since raw postings contain text unrelated to workplace skills, we use only the content after "Job Description:", which primarily describes position requirements and required skills. Then, we match skills in each description with the entries in the standardized taxonomy (case-insensitive) and replaced them with a mask token. To comply with BERT's 512-token limit, we construct balanced context win-

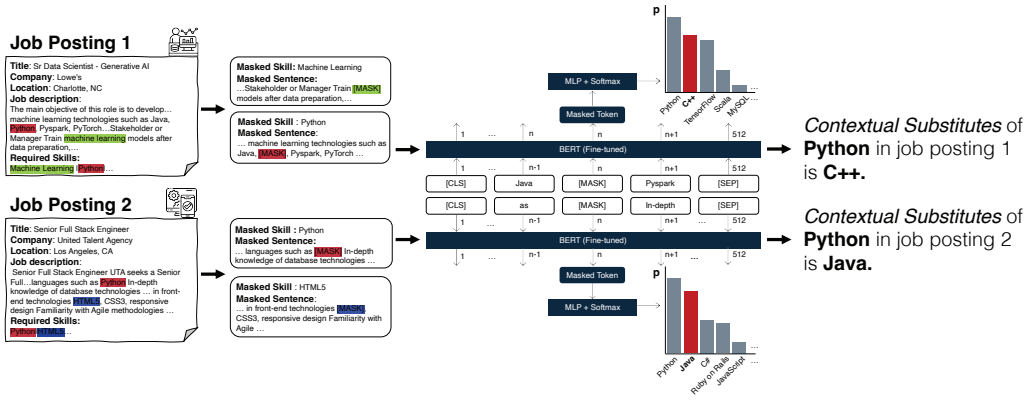


Figure 1: **Overview of the contextual skill substitution framework**. From two separate postings, required skills are extracted and converted into masked sentences, where the target skill is replaced with the [MASK] token while preserving surrounding text. Our model predicts substitutes via an MLP + Softmax layer. We define contextual substitutes as the highest-probability skill after excluding the original masked skill.

dows around each masked skill by iteratively expanding from both sides, using the opposite side when needed to fully utilize the available context length. The resulting masked-skill instances are split into training and test sets, each containing approximately 2.9M valid instances. We fine-tune the model on the training set and use subsets of the held-out test set for validation and full sets for downstream analyses.

2.3 Pre-Training and Fine-Tuning BERT

Our model utilizes a BERT encoder (Devlin et al., 2019) with an MLP and softmax layer to predict masked skills from their surrounding context. For training, we adopt a two-step approach. We first *pre-train* the encoder on domain-specific job postings using 15% random token masking, then *fine-tune* specifically for masked-skill prediction over three epochs using the AdamW optimizer with a maximum sequence length of 512 tokens, a batch size of 64, a learning rate of 2×10^{-5}. Fine-tuning was conducted on a single NVIDIA A100 GPU and took approximately one day. The model minimizes the cross-entropy loss ($\mathcal{L}_{\text{CE}} = -\log(\hat{y}_{s^*})$) where $\hat{y}_{s^*}$ is the predicted probability of the ground-truth skill s^*. This yields a probability distribution over the skill vocabulary $\mathcal{V}$, enabling the model to capture functional substitutes based on surrounding semantic signals.

2.4 Interpreting Contextual Substitution via Sparse Autoencoders

To understand why our model identifies effective contextual substitutes, we use sparse autoencoders (SAEs) to examine how it represents the same skill across industries and time. Standard transformer hidden states are typically polysemantic, entangling multiple contextual signals within a single dense vector; SAEs disentangle them into interpretable, sparsely activated features, isolating the industry- and time-specific signals behind the model's substitution decisions. Concretely, for each layer $l \in \{1, \ldots, L\}$, we extract the contextualized hidden representation at the masked-token position and train an encoder–decoder SAE that projects it into a higher-dimensional sparse feature vector and reconstructs the original representation (Wu et al., 2024). The objective combines a reconstruction loss with an ℓ_1 penalty for sparsity and an ℓ_2 penalty for stability. We freeze the BERT encoder and train each layer-wise SAE for three epochs using AdamW ($m = 4096$, learning rate 1×10^{-4}, $\lambda_{\ell_1} = 10^{-3}$, $\lambda_{\ell_2} = 10^{-4}$).

3 Model validation

To assess the model's capability to infer masked skills and produce reasonable contextual substitutes, we employ two complementary evaluation measures. First, we evaluate masked-skill prediction accuracy to verify that the model can recover the correct skill from its surrounding context. Second, we assess contextual substitution quality using a semantic validation score, which measures whether the substituted sentence preserves the meaning of the original job description. This type of sentence-level evaluation is widely used in lexical substitution (LS) (McCarthy and Navigli, 2007) and smart word suggestion (Rogers et al., 2023; Jabbari et al., 2010; Zhou et al., 2019).

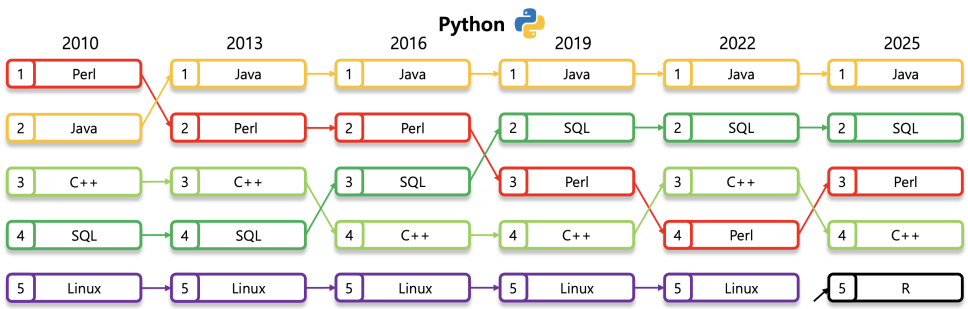


Figure 2: **Longitudinal substitute trajectories of `Python`.** Top-ranked substitutes for `Python` evolve dynamically between 2010 and 2025. This trajectory demonstrates that the model captures time-dependent shifts in skill relationships as technologies and paradigms change. Trajectories of `C`, `JavaScript`, and `Java` are in Appendix A.3

3.1 Model Accuracy

We first assess whether the model has effectively learned domain-specific skill distributions by evaluating its ability to recover masked skills on a held-out test set of 1.4 million samples. For each masked skill s^*, the model outputs a probability distribution $P(s \mid X)$ over the candidate skill vocabulary $\mathcal{V}$ ($|\mathcal{V}| = 2{,}012$), and we compute Top-n accuracy as the fraction of cases in which the ground-truth skill appears among the n highest-probability candidates. The model achieves strong predictive performance, with Top-1/3/5 accuracy of 0.808/0.917/0.944. These results indicate that the model reliably infers masked skills from their surrounding context, learning robust domain-specific representations that support the subsequent analyses of contextual skill substitutability.

3.2 Contextual Substitution Validation

A valid substitute must preserve contextual meaning without redefining job functions or altering a role's technical and organizational context: substituting `Python` with `R` preserves analytical intent, whereas substituting with `Excel` would fundamentally alter the hiring intent. We operationalize this by extending Zhou et al. (2019) with a LLaMA-based contextual validator (Grattafiori et al., 2024) that measures substitution stability while avoiding circularity with our BERT-based model, quantifying the semantic similarity between original and modified postings. Formally, the validation score is defined as:

$$s(X', X, k) = \sum_{i=1}^{L} w_{i,k} \cdot f\big(h(x_i \mid X),\, h(x'_i \mid X')\big),$$

where X is the original sentence and X' is the sentence after replacing the target skill with a candidate substitute. Since many workplace skills are multi-token phrases, we use the first token of each skill span as the anchor position k. The representation $h(x_i \mid X)$ concatenates the final four hidden layers of LLaMA, $f(\cdot,\cdot)$ denotes cosine similarity, and the weight $w_{i,k}$ is the average self-attention from token i to k, aggregated across all layers and heads. Higher scores indicate that the substitution better preserves the original contextual meaning.

3.2.1 Co-occurrence-based Substitutability

As a baseline, given two skills s_1 and s_2, we estimate their substitutability using the conditional probability $P(s_2 \mid s_1)$ derived from their co-occurrence in job postings: if s_2 frequently appears in postings that also mention s_1, it can be regarded as a plausible substitute for s_1. Formally, $P(s_2 \mid s_1) = \frac{\text{count}(s_1, s_2)}{\text{count}(s_1)}$, where $\text{count}(s_1, s_2)$ is the number of postings containing both s_1 and s_2, and $\text{count}(s_1)$ is the total number of postings containing s_1. For each target skill s^*, we rank candidates by $P(s \mid s^*)$ and select the highest-ranked ones as substitutes.

3.2.2 Embedding-based Substitutability

Following previous work (Van-Duyet et al., 2017; Mikolov et al., 2013), we apply Skip-gram with Negative Sampling (*word2vec*) to learn dense skill embeddings, treating the skill requirements in each posting as co-occurring terms and shuffling their order to mitigate positional bias (Murray et al., 2023). For each target skill s^*, we select the highest-cosine-similarity skill as its substitute. While this captures global co-occurrence, it remains context-agnostic, ignoring how interchangeability shifts across job postings.

3.2.3 Contextual Substitutability (our approach)

Our approach leverages the probability distribution produced by the masked-skill prediction model, which assigns a likelihood $P(s \mid X)$ to each candidate skill $s \in \mathcal{V}$ given the surrounding context X. Since the most probable skill typically reconstructs the original masked skill s^*, we instead take the second-highest-probability candidate as the contextual substitute,

$$\hat{s}_{\text{sub}} = \arg \max_{s \in \mathcal{V} \setminus \{s^*\}} P(s \mid X).$$

We compared this choice against alternative selection strategies—top-k, probability-threshold, and temperature-based selection—and found that the second-top candidate is the most plausible substitute (Appendix A.1).

3.3 Comparison with Baselines

Table 1 reports the mean validation scores across models. Our approach achieves the highest score, outperforming both the *word2vec* and conditional probability baselines. Notably, it also consistently surpasses *GPT-5.1*, a cutting-edge large language model, in a comparison over 14,000 samples (prompt detailed in Appendix A.1), which is particularly compelling given that a compact BERT encoder fine-tuned on domain-specific job postings outperforms a massive general-purpose LLM. As an illustrative example, Table 2 shows substitute predictions for `Python` in 2010, when it often functioned interchangeably with other server-side scripting and database languages. Our model retrieves functional substitutes such as `Perl`, `Java`, and `C++`, whereas *word2vec* and conditional probability emphasize related libraries or frequent co-occurrences. Although *GPT-5.1* also produces reasonable alternatives, our compact, domain-adaptive model is more scalable and achieves higher validation scores. The detailed ranking procedure is described in Appendix A.2.

One concern is that these results possibly rely on a LLaMA-based validator, which may introduce evaluator-specific bias. To rule out this possibility, we conduct an additional experiment that re-evaluates all models with two alternative models, Qwen2.5-3B-Instruct (Qwen et al., 2025) and Gemma-2-2B-it (Team et al., 2024). As shown in Appendix Table A.5, our method remains consistently superior across all evaluators.

Table 1: **Mean validation scores.**

Model	Average SV
Baseline Comparisons	
Ours	**0.8118**
word2vec	0.6491
Conditional Probability	0.7996
LLM-based Comparison	
Ours	**0.8143**
GPT-5.1	0.7657

Table 2: **Substitutes of Python (2010) across models.**

Model	Top-3 Substitutes
Ours	Perl, Java, C++
word2vec	Keras, PyTorch, MLOps
Cond. Prob.	SQL, Java, Scripting
GPT-5.1	Bash, Perl, JavaScript

4 Findings

This section summarizes our main findings. First, the model's substitutes agree with the alternatives that HR professionals list as interchangeable, performing close to a human expert (Section 4.1). Second, a single skill maps to different substitutes depending on the job's field and year, showing that substitutability depends on context (Section 4.2). Third, using layer-wise SAEs, we trace where this context-dependent substitution comes from inside the model: deeper layers separate the same skill by industry and year (Section 4.3). Finally, our simulation shows a practical application for employers posting jobs: replacing a required skill with a more popular substitute helps a job posting reach more candidates (Section 4.4).

4.1 Do Predicted Substitutes Match Human Experts?

We find that our model produces skill substitutes that match those of human HR professionals at near-expert agreement. The most direct test of this claim is whether its predicted substitutes align with skill alternatives that employers themselves declare interchangeable. Since no established benchmark exists for contextual skill substitution, we exploit naturally occurring patterns of the form "`{skill A} or {skill B}`" as a data-driven proxy for employer-stated substitutability. Crucially, these postings are authored by HR professionals in real hiring decisions, so the substitutable pairs they

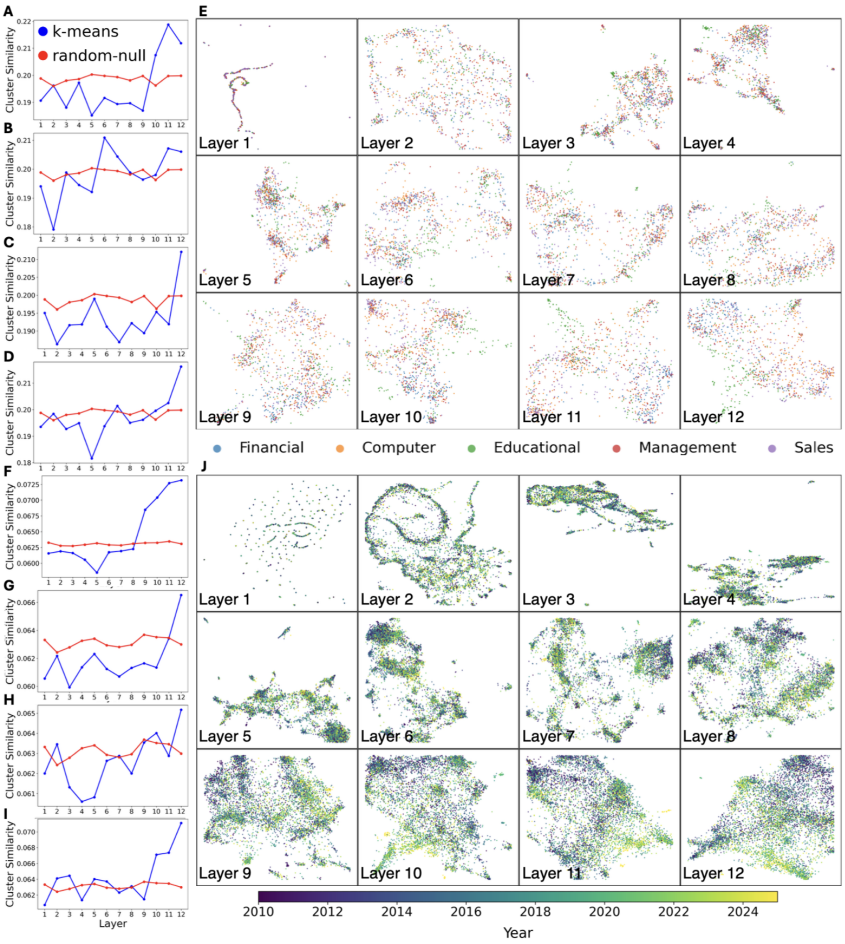


Figure 3: **Layer-wise industry and temporal specialization.** (Left) Element-centric cluster similarity for four programming languages (`Python`, `C`, `Java`, and `JavaScript`), comparing k-means against a random baseline across industries (A-D) and years (F-I). (Right) UMAP visualizations of Python representations using SAE features, colored by industry (E) and year (J). Results demonstrate deeper layers encode specialized, industry-sensitive, and temporally adaptive representations.

specify constitute expert-grounded ground truth rather than a heuristic approximation.

From a 2.9M-instance held-out test set, we extract 17,776 candidate cases matching this pattern and manually filter them to 1,000 high-quality pairs of genuinely substitutable skills. We then mask one side of each pattern, e.g., "`[MASK] or Python`," treating the unmasked skill (`Python`) as context and the masked skill as the ground-truth substitute. Our model and *GPT-5.1* predict the masked skill directly; for the conditional probability and *word2vec* baselines, which lack masked textual context, we instead issue the unmasked skill as a query and compare their top-ranked candidates against the masked ground-truth skill.

Our model achieves Top-1/3/5 accuracy of 0.831/0.957/0.982, substantially outperforming *GPT-5.1* (0.533/0.675/0.730), the conditional baseline (0.356/0.557/0.582), and *word2vec* (0.011/0.013/0.014). The margin is striking: on the most stringent Top-1 setting, our model recovers the exact substitute named by HR professionals in over 83% of cases—approaching expert-level agreement and far surpassing even a cutting-edge general-purpose LLM. This demonstrates that our approach captures context-sensitive substitutability that co-occurrence- and prompt-based methods fundamentally miss.

4.2 Context-Dependent Substitution across Occupations and Time

Substitutability is not fixed but varies systematically along two axes: occupational domain and time. To examine occupational variation, we categorize job postings using the 23 hierarchical SOC Major Groups, a standard taxonomy in labor market analysis (Hershbein and Kahn, 2018; Foote and Ryan, 2015; Azar et al., 2022). Identical

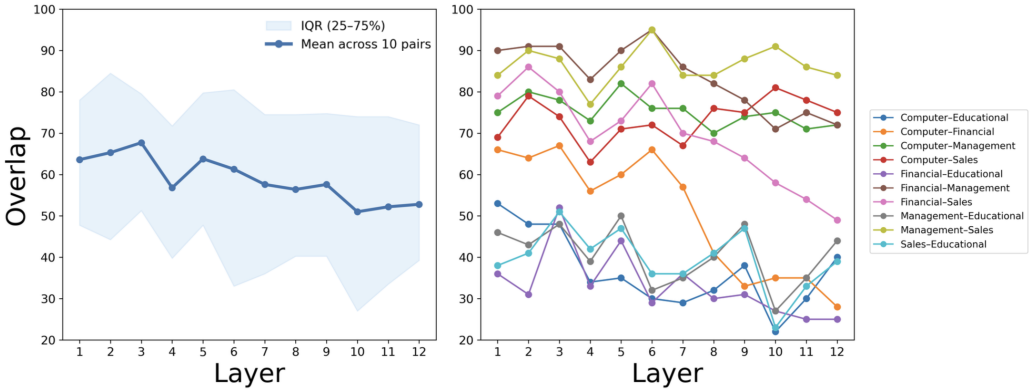


Figure 4: **Layer-wise specialization of SAE features for the `Python` skill.** (Left) Mean overlap of Top-K SAE features across BERT layers, aggregated over all industry pairs. The shaded region denotes the interquartile range (25–75%). (Right) Pairwise feature overlap across layers for each industry pair, illustrating how SAE feature overlap differs by industry as depth increases. Both panels share the same y-axis scale for direct comparison.

skills yield distinct substitutes depending on the domain; for example, `C` is frequently substituted with `Embedded C` in engineering contexts, whereas the model identifies `R` as a more plausible substitute in healthcare research roles—a distinction lost in context-agnostic models. Detailed substitute patterns across domains and SOC abbreviations are provided in Table A.7 and Table A.6.

Substitution is equally dynamic over time. Between 2010 and 2025, substitution patterns shift markedly, reflecting broader technological and occupational change (Fig. A.2). For instance, `Python`'s primary substitutes transition from `Perl` to `SQL` (Fig. 2), `JavaScript` moves from `jQuery` to `TypeScript`, and `C` increasingly aligns with `Embedded C` and `C++` (Fig. A.3). As the surrounding job landscape evolves, the model identifies substitutes that track contemporary technological demands, providing a scalable lens for observing skill evolution and actionable insights for workers and policymakers navigating a shifting labor market.

4.3 How the Model Encodes Context: An SAE Analysis

To examine how the model captures this contextual dependence internally, we analyze how skill representations change across the layers of the BERT encoder using layer-wise SAEs. The idea is that although a given skill token appears in many postings, its meaning varies by industry and over time; if the model encodes this variation, its representations should become more context-specific in deeper layers.

The analysis shows a consistent pattern of increasing contextual specialization across layers. UMAP projections of SAE features show that early-layer representations of skills such as `Python` remain largely mixed across industries, whereas deeper layers form a more structured organization aligned with SOC major groups (Fig. 3E). To quantify this, we cluster SAE feature activations at each layer for `Python`, `C`, `Java`, and `JavaScript` (Fig. 3A–D) and measure their alignment with occupational labels using element-centric cluster similarity (Gates et al., 2019). Across all four skills, similarity increases with depth and exceeds a random baseline around layers 10–12. A similar trend appears along the temporal dimension: aligning SAE-derived clusters with year labels yields similarity scores that also rise in deeper layers (Fig. 3F–J). Together, these findings suggest that skill representations become more specialized along both occupational and temporal dimensions.

A complementary feature-overlap analysis supports this specialization. As depth increases, the set of SAE features jointly activated across industries becomes smaller (Fig. 4), so the representation of a single skill—for example, `Python`—relies on more industry-specific features despite its unchanged surface form. A similar temporal analysis (Appendix A.5) shows the same decline in overlap as the year gap widens. Taken together, these results indicate that the model's contextual substitutions are mirrored by context-differentiated internal representations rather than by surface lexical similarity, consistent with a gradual rather than abrupt change in how each skill is used.

Figure 5: **Market exposure gain across job postings from January to June 2025.**

4.4 Quantifying Market Exposure Gain from Contextual Substitution

Finally, we examine a practical use of our approach: improving the visibility of job postings to potential candidates without changing the employer's hiring intention. To measure this *exposure gain*, we compare the popularity of the original and substituted skills using Google Trends scores. Let x be the trend score of the original posting and y that of the substituted version. The exposure gain, defined as $(y - x)/x$, captures the relative improvement in visibility when a trending substitute replaces the original skill. Since a posting may list several required skills in different contexts, we compute this gain for each contextual instance, where a positive gain means that the substitution increases visibility while keeping the intended occupational context and skill requirements intact.

As shown in Figure 5, many postings achieve higher exposure through contextual substitution without altering their functional requirements. More than half of all postings (approximately 64.09%) contain at least one positive-margin substitution, and some show large relative gains. Because the relative gain $(y - x)/x$ can be unstable when the original trend score x is small, we also evaluate three alternative metrics for robustness—absolute gain $(y - x)$, clipped relative gain $(y - x)/\max(x, c)$, and log-ratio gain $\log((y + \epsilon)/(x + \epsilon))$—and find that the exposure pattern does not depend solely on near-zero baseline scores (Appendix A.3). Overall, these results point to the practical value of contextual substitution for improving recruitment visibility and for understanding labor market dynamics.

5 Conclusion

In this study, we addressed the challenge of identifying contextually appropriate substitutes for workplace skills. Traditional approaches rely on global co-occurrence or static similarity, and thus cannot capture the context-dependent nature of skill substitutability in job postings. We introduced a masked language modeling framework that predicts skill substitutes from the surrounding job description context, giving a more fine-grained view of how skills can be functionally interchangeable across occupations and over time.

Using a large-scale Lightcast dataset of U.S. job postings, we show that our model outperforms context-agnostic baselines—*word2vec*, conditional probability, and *GPT-5.1*—under both semantic validation and explicit substitutable-pair evaluation. Most notably, when tested against skill alternatives that HR professionals mark as interchangeable in real postings, our model recovers these expert-stated substitutes with near-expert accuracy. Our findings also show that substitutability is not static: the same skill can yield different substitutes depending on occupational domain and time period. These results highlight the value of modeling skill relationships as context-sensitive and temporally evolving rather than as fixed global similarities.

We further provide a representational analysis using sparse autoencoders (SAEs), showing that deeper BERT layers align more closely with occupational and temporal distinctions in skill usage. The SAE analysis indicates that the model captures how the same skill is used differently across industries and over time. Finally, our market exposure simulation shows that contextually compatible, trending substitutes can improve the Google Trends-based visibility of job postings, pointing to practical value for employers, workers, and policymakers navigating a rapidly changing labor market.

Overall, our framework offers a scalable, context-aware approach for identifying skill substitutes, tracking the evolution of labor-market demands, and supporting more flexible matching between job requirements and worker capabilities.

6 Limitations

Our study has several limitations. First, semantic validation is based on LLM evaluators. Although results are robust across multiple evaluators, the scores should be viewed as comparative indicators rather than absolute ground truth. Second, our SAE analysis is correlational: it shows that representations become increasingly aligned with industry and temporal context, but it does not establish that specific SAE features causally drive substitution decisions. Third, our exposure analysis relies on Google Trends as an approximate proxy for skill visibility, not as a direct measure of job-posting exposure, applications, interviews, or hiring outcomes. Finally, because our model identifies substitutes from job-posting context, it is important for employers to write postings as specifically as possible. When employers clearly describe their technical environment and role requirements, the model can help HR professionals and job seekers identify plausible alternative skills or candidate profiles. However, vague or underspecified postings may lead to poorly matched substitutes.

References

Ahmad Alabdulkareem, Morgan R Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. 2018. Unpacking the polarization of workplace skills. *Science advances*, 4(7):eaao6030.

Liudmila Alekseeva, José Azar, Mireia Giné, Sampsa Samila, and Bledi Taska. 2021. The demand for ai skills in the labor market. *Labour economics*, 71:102002.

Jim Allen and Rolf Van der Velden. 2001. Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search. *Oxford economic papers*, 53(3):434–452.

Katharine A Anderson. 2017. Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences*, 114(48):12720–12724.

José Azar, Ioana Marinescu, and Marshall Steinbaum. 2022. Labor market concentration. *Journal of Human Resources*, 57(S):S167–S199.

Lex Borghans, Bas Ter Weel, and Bruce A Weinberg. 2014. People skills and the labor-market outcomes of underrepresented groups. *Ilr Review*, 67(2):287–334.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. https://www.anthropic.com/research/towards-monosemanticity. Technical report.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

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

David J Deming and Kadeem Noray. 2020. Earnings dynamics, changing job skills, and stem careers. *The Quarterly Journal of Economics*, 135(4):1965–2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

David Evans, Claire Mason, Haohui Chen, and Andrew Reeson. 2024. Accelerated demand for interpersonal skills in the australian post-pandemic labour market. *Nature Human Behaviour*, 8(1):32–42.

Lucrezia Fanti, Dario Guarascio, and Matteo Tubiana. 2021. Skill mismatch and the dynamics of italian companies' productivity. *Applied Economics*, 53(59):6790–6803.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Luz Adriana Florez and Leidy Gomez. 2024. The impact of skill mismatch on unemployment, informality, and labour turnover. *The Economic and Labour Relations Review*, 35(4):980–999.

Christopher L Foote and Richard W Ryan. 2015. Labor-market polarization over the business cycle. *NBER Macroeconomics annual*, 29(1):371–413.

Alexander J Gates, Ian B Wood, William P Hetrick, and Yong-Yeol Ahn. 2019. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific reports*, 9(1):8574.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Francis Green and Yu Zhu. 2010. Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. *Oxford economic papers*, 62(4):740–763.

Fatih Guvenen, Burhan Kuruscu, Satoshi Tanaka, and David Wiczer. 2020. Multidimensional skill mismatch. *American Economic Journal: Macroeconomics*, 12(1):210–244.

Charles Ham, Rebecca N Hann, MaryJane Rabier, and Wenfeng Wang. 2025. Auditor skill demands and audit quality: Evidence from job postings. *Management Science*, 71(7):5805–5829.

Michael J Handel. 2003. Skills mismatch in the labor market. *Annual Review of Sociology*, 29(1):135–165.

Brad Hershbein and Lisa B Kahn. 2018. Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–1772.

Beth F Ingram and George R Neumann. 2006. The returns to skill. *Labour economics*, 13(1):35–59.

Sanaz Jabbari, Mark Hepple, and Louise Guthrie. 2010. Evaluating lexical substitution: Analysis and new measures. In *LREC*.

Seongwoon Kim, Yong-Yeol Ahn, and Jaehyuk Park. 2024. Labor space: A unifying representation of the labor market via large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2441–2451.

Reiner Kraft, Chi Chao Chang, Farzin Maghoul, and Ravi Kumar. 2006. Searching with context. In *Proceedings of the 15th international conference on World Wide Web*, pages 477–486.

Reiner Kraft, Farzin Maghoul, and Chi Chao Chang. 2005. Y! q: contextual search at the point of inspiration. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 816–823.

Kai Liu, Kjell G Salvanes, and Erik Ø Sørensen. 2016. Good skills in bad times: Cyclical skill mismatch and the long-term effects of graduating in a recession. *European Economic Review*, 84:3–17.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

Seamus McGuinness, Konstantinos Pouliakas, and Paul Redmond. 2018. Skills mismatch: Concepts, measurement and policy approaches. *Journal of Economic Surveys*, 32(4):985–1015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Dakota Murray, Jisung Yoon, Sadamori Kojaku, Rodrigo Costas, Woo-Sung Jung, Staša Milojević, and Yong-Yeol Ahn. 2023. Unsupervised embedding of trajectories captures the latent structure of scientific migration. *Proceedings of the National Academy of Sciences*, 120(52):e2305414120.

Jaehyuk Park, Ian B Wood, Elise Jing, Azadeh Nematzadeh, Souvik Ghosh, Michael D Conover, and Yong-Yeol Ahn. 2019. Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature communications*, 10(1):3449.

Glenda Quintini. 2014. Skills at work: How skills and their use matter in the labour market.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and 1 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Vineeth Rakesh, Suhang Wang, Kai Shu, and Huan Liu. 2019. Linked variational autoencoders for inferring substitutable and supplementary items. In *Proceedings of the twelfth acm international conference on web search and data mining*, pages 438–446.

Ahmed Rashed, Shereen Elsayed, and Lars Schmidt-Thieme. 2022. Context and attribute-aware sequential recommendation via cross-attention. In *Proceedings of the 16th ACM conference on recommender systems*, pages 71–80.

Roland van Rens Rathelot and See-Yu Thijs Chan. 2023. Rethinking the skills gap. *IZA World of Labor*.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Findings of the association for computational linguistics: Acl 2023. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Di Tong, Lingfei Wu, and James Allen Evans. 2021. Low-skilled occupations face the highest upskilling pressure. *arXiv preprint arXiv:2101.11505*.

Le Van-Duyet, Vo Minh Quan, and Dang Quang An. 2017. Skill2vec: Machine learning approach for determining the relevant skills from job description. *arXiv preprint arXiv:1707.09751*.

Etienne Wasmer. 2001. Measuring human capital in the labor market: the supply of experience in 8 oecd countries. *European Economic Review*, 45(4-6):861–874.

Andrew Weaver and Paul Osterman. 2017. Skill demands and mismatch in us manufacturing. *ILR Review*, 70(2):275–307.

John Wu, David Wu, and Jimeng Sun. 2024. Dila: Dictionary label attention for mechanistic interpretability in high-dimensional multi-label medical coding prediction. *arXiv preprint arXiv:2409.10504*.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3368–3373.

A Appendix

A.1 Substitute-selection strategy analysis

The masked-skill prediction model outputs a full probability distribution over candidate skills for each masked instance. Since our downstream analyses require selecting a single contextual substitute, we compare several substitute-selection strategies to justify our use of the highest-probability non-ground truth candidate. Specifically, we evaluate the highest-probability non-original candidate, top-k selection, probability-threshold-based selection, and temperature-based sampling over the top-10 candidates.

For the highest-probability non-ground truth strategy, we remove the ground truth masked skill from the predicted distribution and select the remaining candidate with the highest probability. For top-k selection, we consider candidates within the top $k = 10$ predictions. For threshold-based selection, we retain candidates whose predicted probability exceeds a predefined threshold. We evaluate two thresholds, 0.005 and 0.001. For temperature-based sampling, we sample from the top-10 candidates after applying temperature scaling to the predicted distribution, using temperatures of 1.5 and 2.0.

Each strategy is evaluated using the LLaMA-based validation scorer described in Section 3.2. Starting from the held-out test set, we evaluate the strategies on the common valid subset of 0.85m masked-skill instances for which all compared strategies produce valid substitutions; this common-subset evaluation accounts for cases where threshold-based methods discard instances with no candidate above the threshold. As shown in Table A.1, the *top non-original selection* achieves the highest mean SV score of 0.824, outperforming top-k selection, threshold-based selection, and temperature-based sampling. These results provide empirical support for adopting top non-original selection as the contextual substitute rule in our main analysis.

Table A.1: Comparison of substitute-selection strategies on a sampled validation set of 37K masked-skill instances. Each strategy is evaluated using the LLaMA-based validation scorer.

Selection strategy	Mean SV
Top non-original selection	**0.8242**
Top-k selection ($k = 10$)	0.8190
Threshold-based selection ($\tau = 0.005$)	0.8216
Threshold-based selection ($\tau = 0.001$)	0.8200
Temperature-based sampling ($T = 1.5$, top-10)	0.8143
Temperature-based sampling ($T = 2.0$, top-10)	0.8095

A.2 Ranking Substitution Procedure

The substitutes reported in Fig. 2, Fig. A.3, Table 2, and Table A.7 are computed by aggregating instance-level predictions at the job-posting level. Since our preprocessed data can contain multiple masked instances of the same target skill within a single posting, we weight each selected substitute by the inverse of the number of target-skill occurrences in that posting. Thus, each posting contributes a total weight of one for a given target skill.

Substitutes are then ranked by summing these weights within the relevant aggregation group: year for the temporal analyses and SOC major group for the occupational analyses. For the *GPT-5.1* baseline, which does not provide prediction probabilities, all Top-5 candidates are assigned uniform weights, and the same aggregation procedure is applied.

A.3 Robustness checks for market exposure gain

The main exposure-gain metric $(y - x)/x$ can be sensitive when the original trend score x is small. To address this concern, we examine the distribution of original trend scores and compute alternative gain metrics. The original trend scores are not concentrated near zero overall, with a median value of 21.5. We also evaluate absolute gain $(y - x)$, clipped relative gain $(y - x)/\max(x, c)$ with $c = 10$, and log-ratio gain $\log((y+\epsilon)/(x+\epsilon))$.

Table A.3: Skill categories and the number of unique skills

Category	**# of Skills**
Software Development Tools	307
Other Programming Languages	102
JavaScript and jQuery	93
Operating Systems	110
AI/ML	171
Mobile Development	53
Scripting	36
Databases	167
API	83
Software Development	364
Query Languages	35
Web Design and Development	185
Networking Software	84
Java	107
C and C++	29
iOS Development	26
Cloud Computing	64
Scripting Languages	54

Table A.2: Robustness checks for market exposure gain. Alternative gain definitions yield consistent exposure patterns.

Metric	**Value**
Median original trend score x	21.5
Mean absolute gain $(y - x)$	18.59
Mean clipped relative gain $(y - x)/\max(x, c)$	259.70%
Mean log-ratio gain $\log((y + \epsilon)/(x + \epsilon))$	0.95

These results suggest that the exposure-gain pattern is robust to alternative definitions and is not solely an artifact of small original trend scores. At the same time, Google Trends should be interpreted as a proxy for market exposure rather than a direct measure of application behavior. Since our dataset does not include application-level outcomes, we leave the validation of the relationship between skill search visibility and actual application rates to future work.

Table A.4: Overview of the Lightcast job posting dataset organized into three main categories: metadata, job classification, and skills & certifications.

Category	Description	Column names
Metadata	Core posting attributes include unique identifiers, posting dates, raw titles and body text, employer details, education requirements, compensation information, and geographic location data.	id, posted, expired, url, title_raw, body, company_name, education_levels, salary, location, etc.
Job Classification	Occupational and industry classification using multiple taxonomies (NAICS, ONET, CIP, SOC, Lightcast LOT).	nasics6, onet_2019, cip6, soc_5, lot_career_area, lot_occupation, etc.
Skills & Certifications	Required skills (general, specialized, and software) and certifications extracted from postings.	skills_name, specialized_skills_name, common_skills_name, software_skills_name, certifications_name.

Prompt: You are an expert in labor economics and job skill analysis. You are given a job posting sentence that contains a [MASK] token. Infer the FIVE most appropriate skills to replace [MASK].

HARD CONSTRAINT:
- Choose ONLY from the allowed skill list.
- Do NOT invent new skills.
- Use EXACT spelling from the list.

Allowed skills **[len(all_skills)]**: **[allowed list]**
Output ONLY this format:
skill1|skill2|skill3|skill4|skill5
No extra text.
job posting: **[masked sentence]**

Figure A.1: Prompt used to generate skill substitution outputs for the GPT-5.1 baseline model.

Table A.5: Evaluator robustness results using alternative LLM evaluators. Both baseline and LLM-based comparisons are evaluated on the same datasets as the LLaMA-based evaluation. Scores indicate average SV.

Model	Qwen2.5-3B-Instruct	Gemma-2-2B-it
Baseline Comparisons		
Ours	**0.8827**	**0.8663**
word2vec	0.7854	0.7490
Conditional Probability	0.8789	0.8602
LLM-based Comparison		
Ours	**0.8320**	**0.8982**
GPT-5.1	0.7886	0.8710

Table A.6: SOC Major Groups and Recommended Abbreviations (single-word).

Code	Full Description	Recommended Abbreviation
11	Management Occupations	Management
13	Business and Financial Operations Occupations	Financial
15	Computer and Mathematical Occupations	Computing
17	Architecture and Engineering Occupations	Engineering
19	Life, Physical, and Social Science Occupations	Science
21	Community and Social Service Occupations	Social Service
23	Legal Occupations	Law
25	Educational Instruction and Library Occupations	Education
27	Arts, Design, Entertainment, Sports, and Media Occupations	Arts
29	Healthcare Practitioners and Technical Occupations	Healthcare Practitioners
31	Healthcare Support Occupations	Healthcare Support
33	Protective Service Occupations	Protective
35	Food Preparation and Serving Related Occupations	Food Serving
37	Building and Grounds Cleaning and Maintenance Occupations	Building Maintenance
39	Personal Care and Service Occupations	Personal Care
41	Sales and Related Occupations	Sales
43	Office and Administrative Support Occupations	Office Administration
45	Farming, Fishing, and Forestry Occupations	Agriculture
47	Construction and Extraction Occupations	Construction
49	Installation, Maintenance, and Repair Occupations	Maintenance Repair
51	Production Occupations	Production
53	Transportation and Material Moving Occupations	Transport
55	Military Specific Occupations	Military

Table A.7: Top substitutes for skills aggregated by SOC Major Group

Truth	Substitution	Major Group
C	R	Management; Financial; Social Service; Healthcare Practitioners; Healthcare Support; Protective; Building Maintenance; Sales; Office Administration; Construction; Production
	Embedded C	Computing; Engineering; Arts; Food Serving; Personal Care; Agriculture; Maintenance Repair; Transport
	C++	Law; Education; Military
	Perl	Science
Python	SQL	Management; Financial; Science; Social Service; Healthcare Practitioners; Protective; Sales; Office Administration; Construction; Military
	Java	Computing; Law; Education; Agriculture; Production
	Perl	Food Serving; Personal Care; Maintenance Repair; Transport
	C++	Engineering; Arts
	R	Healthcare Support
	DevOps	Building Maintenance
Java	Python	Management; Engineering; Science; Education; Personal Care; Sales; Office Administration; Construction; Production; Transport
	JavaScript	Social Service; Arts; Healthcare Support; Maintenance Repair
	C++	Healthcare Practitioners; Protective; Military
	DevOps	Law; Building Maintenance
	Java 8	Computing; Agriculture
	SQL	Financial
	Linux	Food Serving
JavaScript	jQuery	Management; Financial; Computing; Arts; Protective; Maintenance Repair; Production; Transport
	Java	Engineering; Science; Social Service; Law; Agriculture; Construction
	HTML5	Education; Personal Care; Sales; Office Administration
	Scripting	Healthcare Practitioners; Food Serving
	JScript	Healthcare Support; Building Maintenance
	Apache Camel	Military

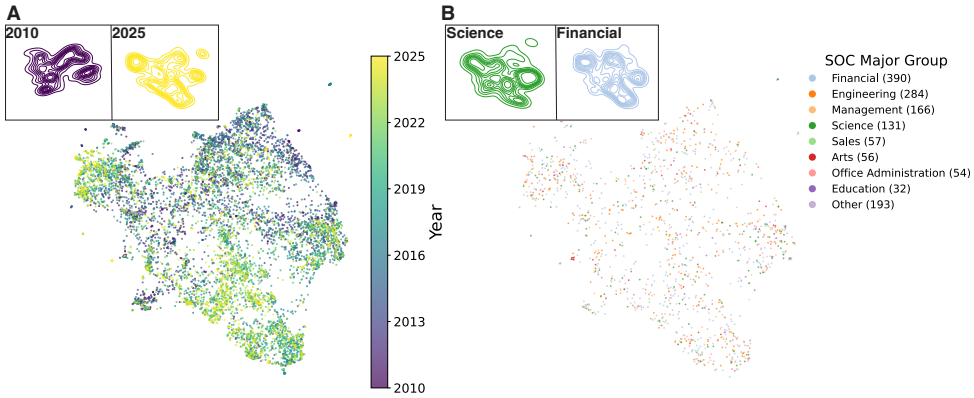


Figure A.2: **Overview of the embedding space for skill `Python`, colored by temporal and SOC major group** We visualize how skill `Python` representations evolve (A) over time (B) and across SOC major groups. For panel (B), job postings in computing occupations are omitted, as they dominate the embedding space. The density maps show that both temporal progression and occupational context shape the organization of the skill embedding space in distinct ways.

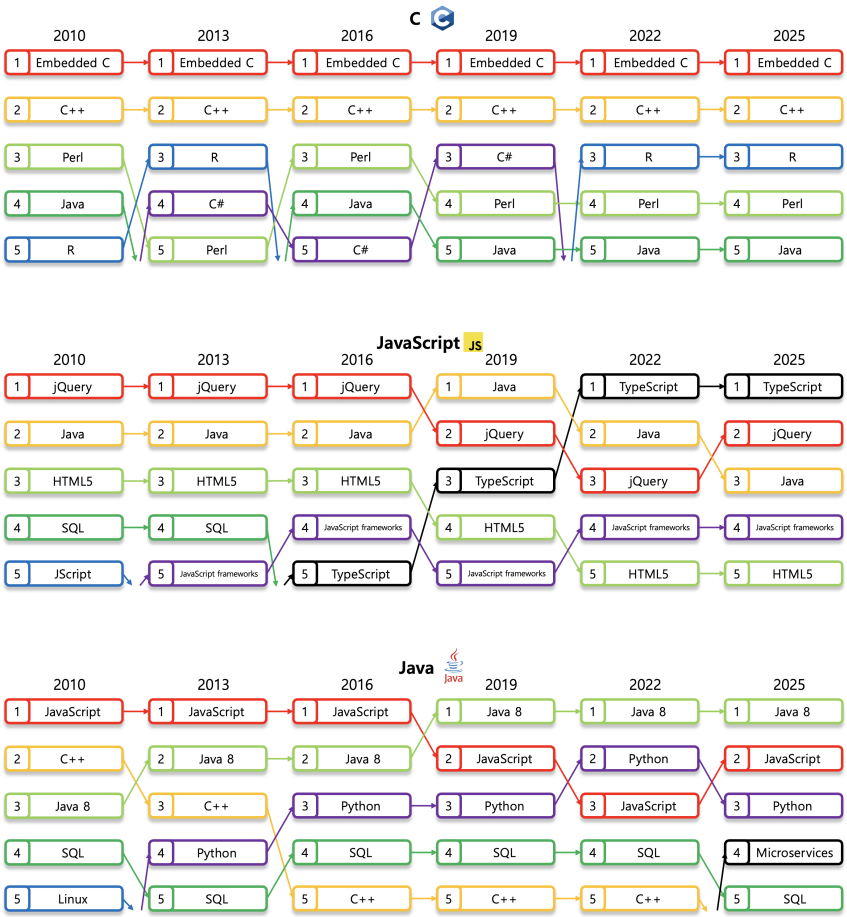


Figure A.3: **Year-wise top–5 substitute trajectories for additional programming skills.**

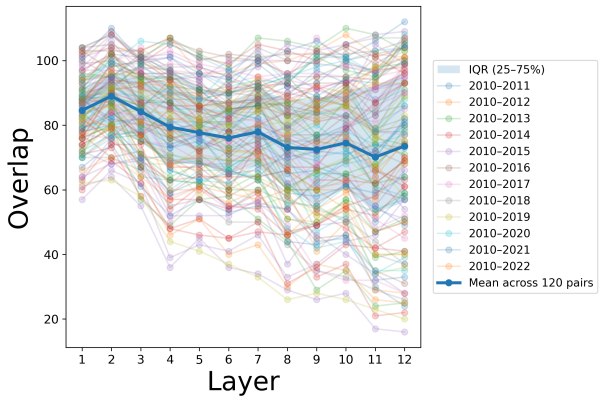


Figure A.4: Layer-wise overlap of highly activated SAE features across years. For each layer, we compute the overlap of Top-K SAE features between pairs of years, where features are ranked by mean activation. Overlap decreases in deeper layers, indicating progressively more temporally specialized representations.

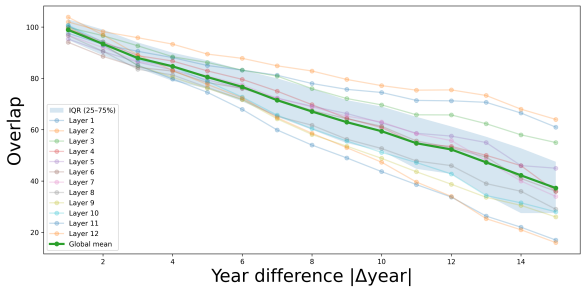
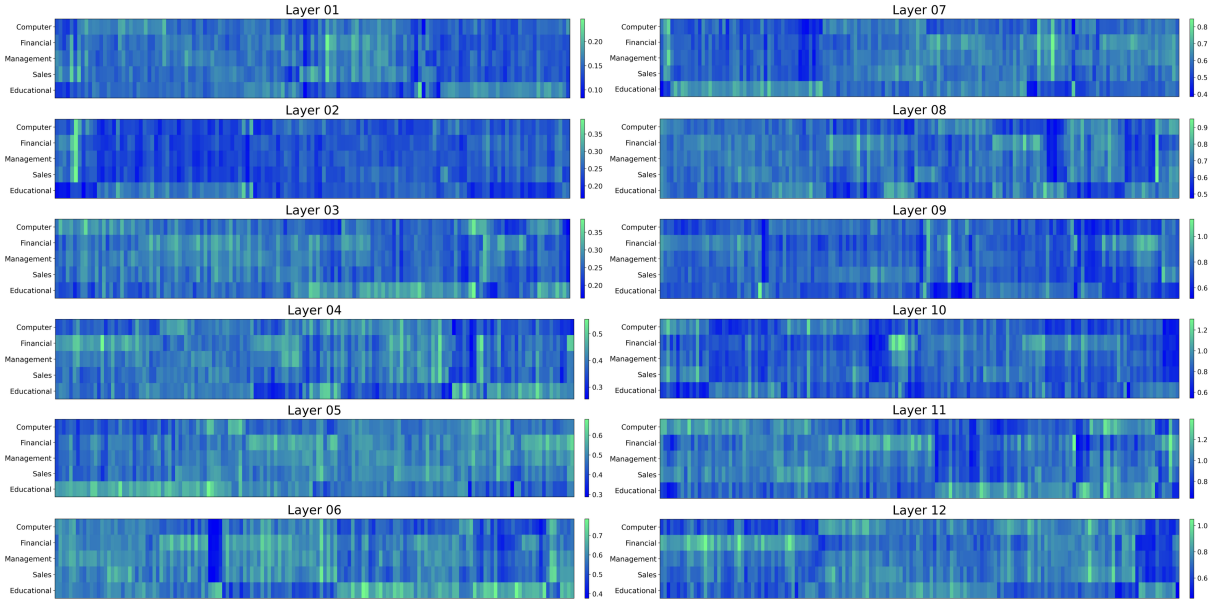
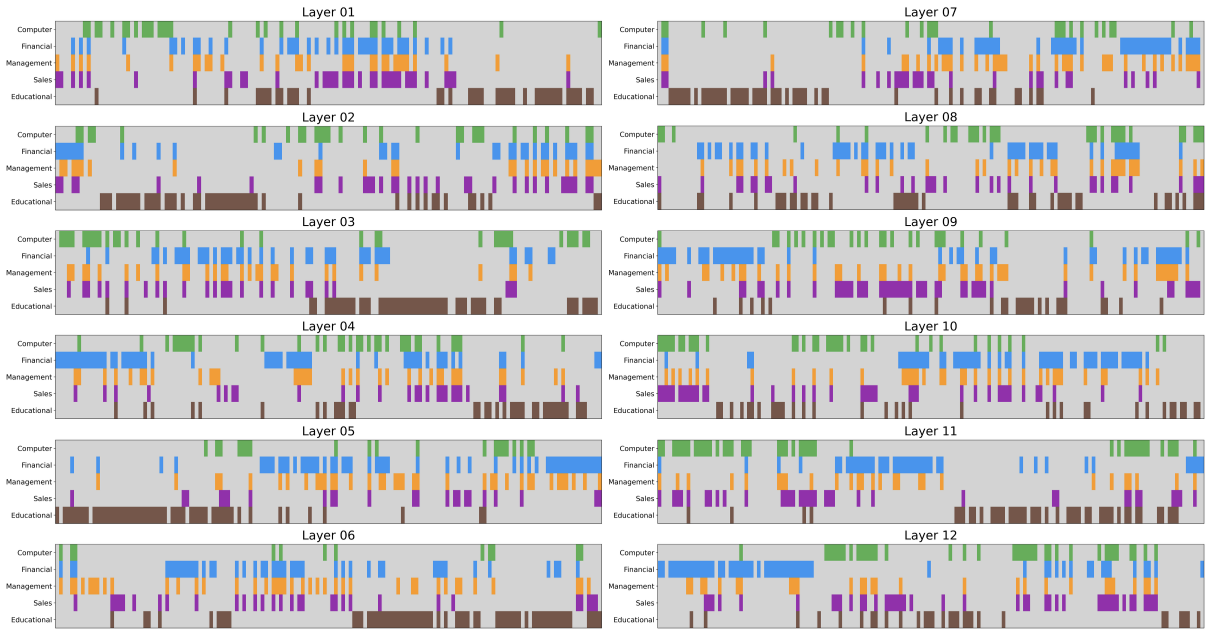


Figure A.5: **Feature overlap vs. year difference $|\Delta\text{year}|$ for the `Python` skill.** Thin lines represent individual layers and the thick line shows the global mean.



(a) Heatmaps of mean SAE activations for the same set of selected features.



(b) Features are binarized by a layer-wise top-$q\%$ threshold (active vs. inactive).

Figure A.6: **Layer-wise SAE activation patterns across industries.** For each layer, we compute the industry-specific mean activation vector and select Top-K SAE features per industry. To emphasize industry-specific signals, we iteratively remove features that appear in the Top-K lists of *all* industries and refill each list to maintain size K (same selection procedure as Figure 5). We then take the union of selected features and visualize (top) the corresponding mean activation values and (bottom) a binary activation map using a layer-wise top-$q\%$ threshold. For readability, union feature columns are reordered by clustering feature activation profiles across industries using agglomerative clustering.

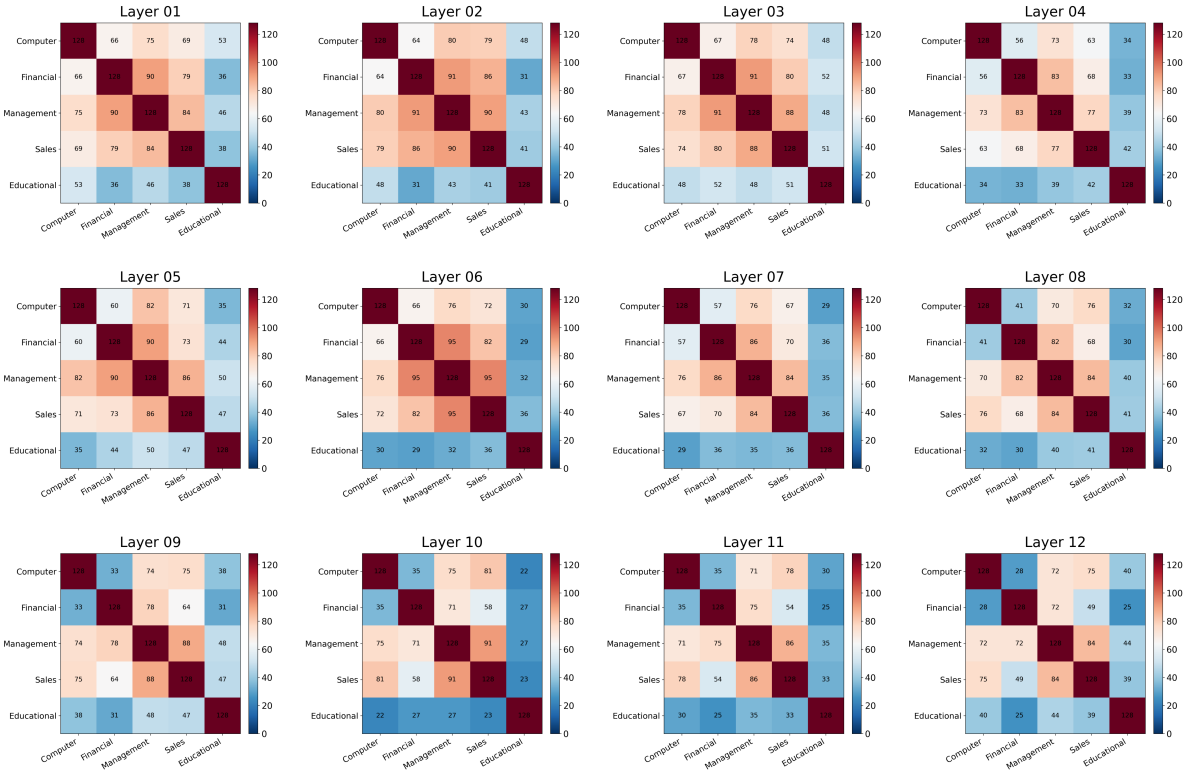


Figure A.7: **Layer-wise overlap heatmaps of SAE features across industries.** Each panel corresponds to one transformer layer and visualizes the number of overlapping top-K SAE features between pairs of industries. The Top-K features are selected using the same iterative procedure as in Figure 4, which removes features shared across all industries to emphasize industry-discriminative activations. Consistent with Figure 4, the magnitude of cross-industry overlap decreases in deeper layers, indicating progressive specialization of representations.

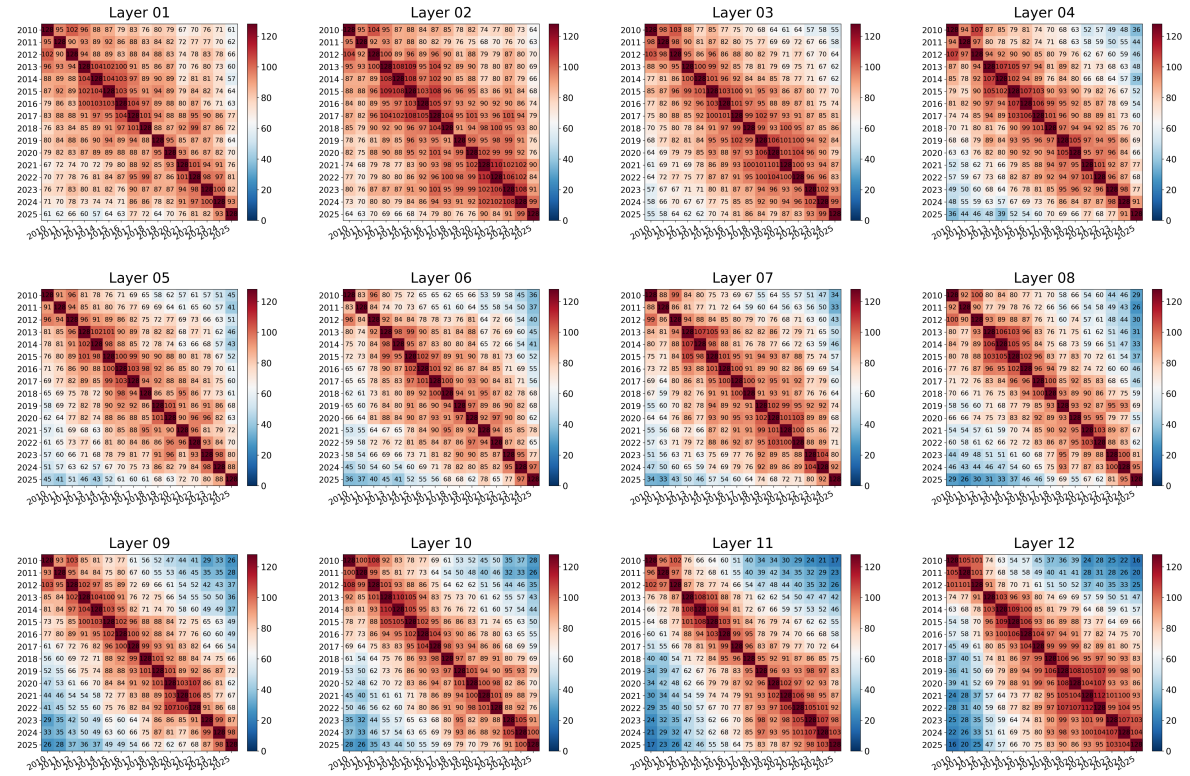


Figure A.8: **Layer-wise overlap heatmaps of SAE features across years.**