SKILL-RAG: RETRIEVAL-AUGMENTED LLMS FOR EXTREME MULTI-LABEL SKILL CLASSIFICATION OF UNSTRUCTURED JOB AD TEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Skill Classification, the process of identifying skill-related information from job ads and mapping it to a fine-grained set of skill labels, opens up new possibilities for research on labor market dynamics. Mainstream skill taxonomies are generally large-scale, such as the ESCO framework with 13,939 labels, making it an extreme multi-label classification (XMLC) task. While Large Language Models (LLMs) prompted with job ads and skill ontologies offer robust automated classification capabilities, deploying them on massive, fine-grained taxonomies introduces excessive inputs, thereby triggering issues of context window constraints, long-context performance degradation, and positional bias. To overcome these challenges, we propose Skill-RAG, a novel pipeline that integrates a Retrieval-Augmented Generation (RAG) framework into the inference process. Specifically, we systematically explore various knowledge base construction paradigms, query granularity strategies (raw vs. span-based retrieval), and multi-stage retrieval pipelines to optimize retrieval accuracy. For downstream LLM-based classification, we comprehensively evaluate prompting and reasoning strategies across three core dimensions: label pool size, label ordering, and thought-guided reasoning. Empirical evaluations confirm that each module plays a distinct role in overall performance, and our proposed pipeline substantially outperforms previous state-of-the-art methods on standard benchmarks.

1 INTRODUCTION

Skill Extraction (SE), the process of identifying and extracting skill-related information from job ads, opens up new possibilities for social science research on labor market dynamics (Deming & Kahn, 2018; Blair et al., 2025). To enable the construction of robust and standardized variables for diverse inference and prediction modeling, Skill Extraction has recently evolved from a tagging task, aiming to identify skill spans within job texts (Zhao et al., 2015; Khaouja et al., 2021; Zhang et al., 2022) into an extreme multi-label classification (XMLC) task, where job ads are linked to a fine-grained set of skill labels from established skill ontologies (Bhola et al., 2020; Decorte et al., 2023; Vermeer et al., 2022). However, skill classification tasks typically leverage large skill ontologies, for instance, the widely used European Skills, Competences, Qualifications and Occupations (ESCO) framework, which encompasses 13,939 skill labels and creates a significant barrier for computational social scientists to explore into the label market dynamics.

Previous approaches to extreme multi-label skill classification, such as BERT-based pipelines (Liu et al., 2017; Chang et al., 2020), are typically constrained by the lack of high-quality, large-scale, and well-balanced manually annotated data for effective model fine-tuning. To mitigate this challenge, alternative approaches have leveraged large language models (LLMs) prompted with job advertisements and skill ontologies for automated classification, attaining robust performance with substantially smaller fine-tuning datasets and minimal task-specific setup (Wang et al., 2025). As the granularity of analysis increases, the scale of skill ontologies expands beyond 10,000 labels, introducing new bottlenecks. The influx of oversized skill taxonomies into LLMs often induces attention dilution and performance degradation. Furthermore, it may become entirely impracticable when the input size surpasses the model's context window (An et al., 2024; Peysakhovich & Lerer, 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we propose **Skill-RAG**, a novel pipeline designed to address the challenges of excessive input and label overload and in extreme multi-label skill classification by integrating a Retrieval-Augmented Generation (RAG) framework into the inference process. We systematically optimize and evaluate strategies across the entire pipeline on the **SkillSpan** dataset, the most widely recognized benchmark in skill classification.

Our framework consists of two main stages:

1) Multi-Stage Span-Based Retrieval: To reduce the candidate label space while maintaining recall, we use a two-stage retrieval setup with dense retrieval in the first pass and a cross-encoder reranker in the second pass. Since the mainstream skill ontologies (e.g., ESCO) typically provide terse and abstract label definitions, we use LLM-generated synthetic text to expand labels with detailed skill descriptions and contexts. Additionally, we introduce span-based retrieval to distill core contents, leveraging a lightweight local LLM to extract skill-related text spans from the raw input before vector retrieval.

2) Downstream LLM Classification: Our pipeline is built on representative models including *Gemini-2.5-Flash* and *DeepSeek-V4.1-Flash*. To mitigate long-context degradation triggered by attention dilution during candidate overload, as well as positional sensitivity issues, we systematically investigate how prompt configuration and reasoning paradigms govern classification accuracy. Specifically, on the prompt level, we examine how input label pool size and label ordering affect model performance. On the reasoning level, considering that the recent shift toward models with native reasoning capabilities (versus traditional non-reasoning generation) may fundamentally alter prompt effectiveness, we compare Chain-of-Thought (CoT) prompting against traditional zero-shot and few-shot baselines across both modes.

Our experiments reveal several practical findings:

External Knowledge Base Construction and Retrieval Granularity: Detailed synthetic skill descriptions significantly boost retrieval precision, whereas uncurated synthetic corpora (e.g., skill-related job ad examples) introduces noise and fail to provide significant gains. Retrieving with extracted skill spans consistently outperforms using the full raw text query, as it removes non-essential context before search.

Label Space Bottleneck: In classification, performance exhibits a clear *inflection point*. Beyond this threshold, the degradation from excessive input and attention dilution outweighs the benefit of capturing more true-positive labels in the candidate pool.

Label Position Bias: Labels located in the middle of prompts are severely neglected due to position bias. Positioning gold labels toward the boundaries (front or back) via reranking effectively increase precision.

Reasoning Modes and Over-Reasoning: Under non-native reasoning generation modes, CoT outperforms few-shot prompting (with zero-shot yielding the lowest accuracy). Crucially, under native reasoning modes, explicit CoT prompts induce an over-reasoning effect that counterproductively harms final classification accuracy.

Our main contributions are as follows: 1. We present **Skill-RAG**, an end-to-end retrieval and classification pipeline that enables scalable extreme multi-label skill extraction while outperforming existing baselines on the SkillSpan benchmark. 2. For the retrieval stage, we systematically analyze the impact of knowledge base construction paradigms and query granularity (raw text vs. extracted spans). 3. For downstream LLM classification, we conduct a systematic empirical analysis on the limits of label context size, label order sensitivity, and the interaction between prompting techniques and native reasoning models.

2 Related Work

2.1 Multi-label Text Classification: From Deep Learning to LLM Reasoning

Extracting skill information from job postings and resumes has been studied for over a decade, initially as sequence tagging and entity matching over unstructured text (Calanca et al., 2019; Gugnani & Misra, 2020; Khaouja et al., 2021; Zhao et al., 2015). These systems locate skill mentions but

rarely normalize them against a shared ontology, which restricts their value for quantitative labor market analysis. The arrival of benchmarks such as SkillSpan (Zhang et al., 2022), together with the adoption of extreme multi-label text classification (XMLC) as the standard framing (Chalkidis et al., 2019), shifted the goal toward assigning labels from taxonomies that contain thousands of concepts.

Deep learning dominated this reframed task for years. Convolutional encoders gave way to label tree based attention models (Liu et al., 2017; You et al., 2019), and pretrained transformers fine-tuned with semantic label indexing and shortlisting became the reference architecture (Chang et al., 2020; Jiang et al., 2021). The training cost of transformers on extreme label spaces was then reduced through recursive multi-resolution fine-tuning (Zhang et al., 2021) and end-to-end multi-resolution training (Kharbanda et al., 2022). Since real-world label distributions are heavily skewed, long-tailed imbalance continues to erode accuracy even for strong baselines (Huang et al., 2021). Within the skill domain, this lineage covers BERT based XMLC over job descriptions (Bhola et al., 2020), CNN classifiers for resumes (Jiechieu & Tsopze, 2021), graph neural networks over skill hierarchies (Goyal et al., 2023), and domain-adapted pretrained models such as RecruitPro (Fang et al., 2023). However, their reliance on large annotated corpora for training remains a primary bottleneck.

Generative LLMs has since become a competitive alternative to fine-tuning for text classification (Fields et al., 2024; Sun et al., 2023b). For skill classification specifically, most systems still keep the LLM away from the final decision over the ontology. SkillGPT prompts an LLM to summarize postings and then retrieves candidate skills by embedding similarity (Li et al., 2023), SkiLLMo couples a fine-tuned segmenter with LLM span extraction followed by vector matching (Malandri et al., 2025), and zero-shot matchers present only a small filtered candidate set to the LLM (Clavié & Soulié, 2023). Parallel work generates synthetic postings with LLMs for training or matching (Decorte et al., 2023; Magron et al., 2024) and adapts general purpose LLMs for extraction with light supervision (Herandi et al., 2024; Nguyen et al., 2024).

2.2 Retrieval-Augmented Generation for Classification

RAG couples a retriever with a generator so that generation is conditioned on external evidence (Lewis et al., 2020), a paradigm consolidated by recent surveys (Fan et al., 2024; Gao et al., 2023). Its retrieval stage has evolved from sparse lexical matching (Robertson & Zaragoza, 2009) toward dense dual encoders trained with contrastive objectives (Izacard et al., 2022; Karpukhin et al., 2020; Reimers & Gurevych, 2019), which scale to millions of chunks but approximate relevance coarsely. Reranking supplies the complementary precision. Cross-encoders and sequence-to-sequence rankers score a query and each candidate jointly (Nogueira & Cho, 2019; Nogueira et al., 2020), late interaction models such as ColBERT trade a modest index overhead for finer token level matching (Khattab & Zaharia, 2020), and LLMs themselves can serve as reranking agents (Sun et al., 2023a).

Retrieval also changes how classification with LLMs is organized. Retrieved demonstrations improve in-context learning (Rubin et al., 2022), and for long inputs a small retrieved evidence set can outperform stuffing the full text into the context window (Xu et al., 2024). The approach is not risk free, since irrelevant chunks measurably erode downstream accuracy (Cuconasu et al., 2024). Closest to our setting, D'Oosterlinck et al. (2024a) propose an infer-retrieve-rank pipeline that makes in-context learning viable over extreme label spaces, and Shang et al. (2026) report gains from retrieval augmentation in multi-label classification of academic texts. What remains open is how retrieval scale, reranking, and LLM reasoning interact when the label space exceeds ten thousand entries and each document carries several true labels, which is exactly the regime our pipeline targets.

2.3 Chain-of-Thought Reasoning for Classification

Chain-of-Thought (CoT) prompting elicits intermediate reasoning steps and produces large gains on arithmetic, symbolic, and commonsense problems (Wei et al., 2022). Follow-up work removed the need for hand-written exemplars (Kojima et al., 2022), aggregated multiple sampled chains by voting (Wang et al., 2023), automated exemplar construction (Zhang et al., 2023), and searched over branching reasoning paths (Yao et al., 2023). The benefits are far from uniform though. A broad replication across text benchmarks finds that CoT helps mainly on math and symbolic reasoning and frequently hurts elsewhere (Sprague et al., 2025), and controlled experiments show that deliberation degrades accuracy on tasks where thinking also makes humans worse, several of which are classification tasks (Liu et al., 2025). Overthinking in long reasoning models wastes tokens and

can invert accuracy on simple inputs (Chen et al., 2025; Sui et al., 2025), while accuracy collapses once problem complexity crosses a critical threshold (Shojaee et al., 2025). Even when the answer improves, the verbalized chain may not reflect what actually drove the prediction (Lanham et al., 2023; Turpin et al., 2023), and recent evidence frames CoT as pattern interpolation bounded by the training distribution rather than genuine inference (Zhao et al., 2026). We read this literature as a warning against treating reasoning traces as a free upgrade for classification. In our setting, CoT pays off only after the candidate label set has been made compact, which suggests that reasoning and retrieval act as complements rather than substitutes.

2.4 Label Overload, Attention Dilution, and Positional Bias

Extreme multi-label classification imposes severe context constraints on LLMs. As context expands, attention spreads thin and recency bias emerges (Peysakhovich & Lerer, 2023), driving steady accuracy degradation even when inputs are padded with neutral tokens (Levy et al., 2024). Benchmarks confirm that while models pass simple needle-in-a-haystack tests, their accuracy drops significantly as sequence length and task complexity grow, failing well before reaching advertised context limits (Hsieh et al., 2024a). Without lexical overlap, performance across most frontier models drops by over 50% at 32K tokens (Modarressi et al., 2025). Across broader long-context suites ($\geq$100K tokens), effective context length consistently falls far short of nominal context windows (Yen et al., 2025; Zhang et al., 2024; An et al., 2025; Kuratov et al., 2024).

Scaling in-context examples hits a similar ceiling for prompt-based classification. Performance degrades or saturates when demonstrations scale into the hundreds (Bertsch et al., 2025; Li et al., 2025; Agarwal et al., 2024). Multi-label classification compounds this issue, as LLMs tend to suppress secondary labels at each step (Ma et al., 2025), reflecting a limited capacity to distribute probability mass over large candidate sets. Passing a ten-thousand-label ontology to an LLM is thus infeasible and counterproductive, as candidates compete for a finite attention budget. Consequently, candidate pre-selection is a fundamental prerequisite for extreme-scale classification rather than a mere efficiency trick—a premise our experiments directly confirm.

Moreover, LLMs process long contexts non-uniformly, exhibiting a U-shaped accuracy curve where middle content is frequently ignored (Liu et al., 2024). This positional sensitivity severely affects selection tasks: permuting candidate choices or demonstration orders causes wide performance swings (Zheng et al., 2024; Wang et al., 2024; Lu et al., 2022), driven by recency and majority biases (Zhao et al., 2021) and non-uniform positional attention weights (Hsieh et al., 2024b). In extreme multi-label classification with hundreds of candidates, where a target label appears in the prompt can dictate success almost as much as whether it was retrieved—an effect we systematically quantify.

3 Proposed Methods

Let $\mathcal{T}$ denote the input text space (e.g., job descriptions) and $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_M\}$ represent a large-scale skill taxonomy containing M discrete skill labels, where $M \gg 10^3$. Given an input text $t_i \in \mathcal{T}$, the goal of extreme multi-label skill classification is to assign a binary ground-truth target vector $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,M}]^\top \in \{0,1\}^M$, where $y_{i,m} = 1$ indicates that the m-th skill label ℓ_m is relevant to t_i, and 0 otherwise.

To address the challenges posed by label overload, we adopt Retrieval-Augmented Generation (RAG), a framework designed to retrieve contextually relevant information from external data sources to enhance downstream LLM inference. To align with the inherent nature of skill classification, we adapt the standard RAG paradigm into **Skill-RAG**.

Specifically, we construct an external knowledge base by combining skill labels with their enriched contexts or descriptions, partitioning the corpus using individual skill labels as discrete chunking units. Rather than performing a brute-force global classification, candidate label chunks and input queries are embedded into a unified vector space. A two-stage pipeline—combining dense retrieval with a cross-encoder re-ranker is then deployed to filter out candidate labels, significantly reducing input cardinality for the downstream LLM while preserving high label recall. Finally, the retrieved candidate labels, alongside the processed query, are concatenated into an optimized prompt for final LLM-based classification. Our proposed architecture is illustrated in Figure 1.

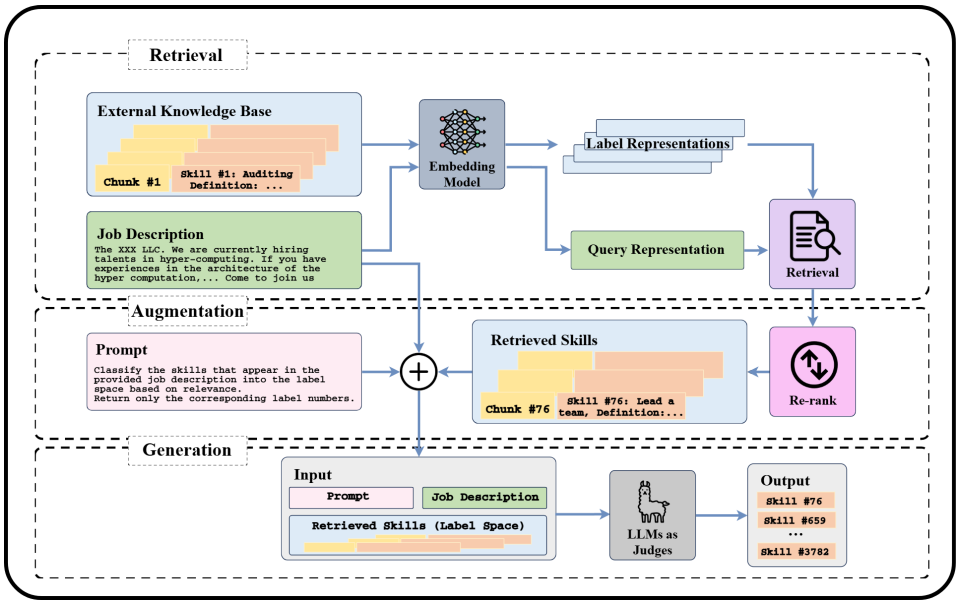


Figure 1: The Skill-RAG Pipeline.

To tailor this framework for skill classification while mitigating the performance degradation and positional bias induced by excessive inputs, we propose several targeted advancements to all steps within the proposed pipeline. These adaptations are elaborated sequentially below.

External Knowledge Base Construction: In classification tasks, external knowledge bases are primarily built upon label-related textual corpora, leveraging label names, definitions, or detailed descriptions to capture the intrinsic semantics of each category (Mueller et al., 2022). While certain standardized ontologies, such as ESCO, inherently provide rich definitional metadata, many other domains lack such comprehensive documentation. To bridge this semantic gap, recent exploratory efforts have turned to label semantic expansion. Specifically, leveraging LLMs for data bootstrapping by synthetically generating label definitions and illustrative instances has emerged as a widely adopted paradigm Clavié & Soulié (2023). However, the empirical efficacy of such synthetic expansion remains an open question, which this study aims to systematically investigate. To operationalize this knowledge base within our retrieval framework, the resulting text corpus is partitioned into discrete chunks at the granularity of individual skill labels. Subsequently, these knowledge units are vectorized into a dense embedding space to enable efficient downstream vector search.

Retrieval Strategies: Prominent paradigms in this domain primarily encompass vector similarity-based Dense Retrieval (DR) and Cross-Encoder Reranking (Busolin et al., 2025). While the former offers exceptional computational efficiency yet suffers from limited semantic alignment precision due to its coarse-grained document embeddings, the latter excels in capturing fine-grained, token-level token interactions at the expense of prohibitive computational consumption (Zhao et al., 2024). To bridge these paradigms, this study synthesizes both approaches into a unified, two-stage retrieval module to optimize classification performance while remaining computationally sustainable. Furthermore, we introduce span-based retrieval by leveraging a lightweight local-deployed LLM to extract skill-related text spans from raw inputs prior to vector search (Wang et al., 2026). This filtering step eliminates non-essential textual noise, preserving only core skill-relevant context to enhance retrieval precision. Through these design choices, we systematically evaluate and compare the efficacy of each retrieval strategy in maximizing recall and mitigating positional bias.

Generation Strategies: Our generation strategies primarily pivot on two core dimensions: prompt configurations and reasoning paradigms. Observing that candidate label pool size and label ordering heavily influence model output, we dynamically vary input label set size and order within the prompt to explore the optimal prompt designs. Regarding reasoning strategies, we contrast Chain-of-Thought (CoT) prompting against standard zero-shot and few-shot baselines to assess the efficacy of explicit reasoning paths. Furthermore, recent commercial LLMs (e.g., Gemini and DeepSeek)

have introduced native reasoning modes. However, such reasoning processes may induce overthinking and semantic drift, ultimately deteriorating classification accuracy (Guo et al., 2026). We systematically evaluate both non-reasoning and native reasoning paradigms to establish the optimal configuration for LLM-based classification.

To sum up, we developed a Retrieval-Augmented Generation (RAG) framework to solve the extreme multi-label skill classification. Specifically, we define a two-stage classification function $f(t_i, \ell_m)$:

$$f(t_i, \ell_m) = g_{\text{LLM}}\left(\ell_m \mid \mathcal{R}_K(s(t_i), \mathcal{K})\right) \tag{1}$$

where $s(\cdot)$ represents a span extraction operator that distills skill-related text spans from raw input t_i, $\mathcal{K} = \{k_m\}_{m=1}^{M}$ denotes an LLM-synthesized external knowledge base enriching each skill label ℓ_m, $\mathcal{R}_K(\cdot)$ denotes the two-stage retrieval operation yielding the top-K candidate label set ($K \ll M$), and $g_{\text{LLM}}(\cdot)$ is the downstream LLM scorer over the candidate pool.

4 Experimental Setup and Results

This study systematically evaluates the empirical performance of these methodologies within the context of extreme multi-label skill classification, with the ultimate goal of formulating a highly effective, optimized pipeline.

4.1 Benchmark and Metrics

This study utilizes SkillSpan Zhang et al. (2022); Decorte et al. (2022), a widely recognized job ad-to-skill benchmark dataset that is formally aligned with the ESCO taxonomy. Given that our proposed architecture encompasses a multi-stage pipeline, we conduct a rigorous component-wise evaluation, which necessitates a comprehensive suite of evaluation metrics. Specifically, we deploy R-Precision@K Chalkidis et al. (2019) to measure the precision-centric alignment of top-ranked candidates, MRR (Mean Reciprocal Rank) to assess the position-aware sensitivity in retrieving the primary target, Recall@K measures the model's ability to capture the complete set of ground-truth labels within the top-K retrieved candidates. This granular metric configuration ensures a holistic validation of our prompting and retrieval strategy.

$$\text{R-Precision@}K = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{|\mathcal{Y}_i \cap \hat{\mathcal{Y}}_i^{(\min(K,|\mathcal{Y}_i|))}|}{\min(K, |\mathcal{Y}_i|)} \tag{2}$$

$$\text{Recall@}K = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{|\mathcal{Y}_i \cap \hat{\mathcal{Y}}_i^{(K)}|}{|\mathcal{Y}_i|} \tag{3}$$

$$\text{MRR} = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{\text{rank}_i} \tag{4}$$

Where N denotes the total evaluation instances, $\mathcal{Y}_i$ and $\hat{\mathcal{Y}}_i^{(K)}$ represent the ground-truth and top-K predicted label sets for instance i, and rank_i is the 1-based position of the first correctly retrieved skill label (set to 0 if no true label is retrieved).

4.2 Reranking and Dense Retrieval

For reranking, we utilize `BAAI/bge-reranker-v2-m3`, a state-of-the-art cross-encoder, to capture deep contextual interactions between the input contexts and candidate skill descriptions. For dense retrieval, we employ `mxbai-embed-large-v1`, a high-capacity Transformer-based embedding model, to map extracted skill spans and label descriptions into a shared dense vector space for inner-product similarity search.

In this experiment, we first compare the performance of cross-encoder reranking with dense retrieval to verify whether reranking outperforms standard retrieval, thereby further pruning the candidate label space while maintaining a high recall rate. Concurrently, we will explore strategies such as

span-based retrieval and label semantic expansion in external knowledge base construction, examining their effects on the final performance.

To compare the performance of cross-encoder reranking against dense retrieval and quantify the effectiveness of the aforementioned strategies, we construct a controlled experimental setup. Specifically, we restrict each candidate pool to a fixed size of 100 labels, consisting of all the gold labels, 50 hard negatives randomly selected from the most semantically similar false labels, and the remaining capacity filled with easy negatives sampled randomly from the broader skill ontology.

For knowledge base construction, we employ few-shot LLM prompting for label semantic expansion, synthesizing detailed skill descriptions and representative job-ad sample sentences to enrich our external knowledge base. The generated text corpus is then partitioned into discrete chunks at the granularity of individual skill labels. The exact prompt templates are documented in Figure A.1 of the Appendix. Consequently, we construct three distinct knowledge base variants: (1) the baseline *ESCO Original Skill Definitions*, (2) the *LLM-Synthesized Skill Descriptions*, and (3) the *LLM-Synthesized Job-Ad Sample Sentences* with varying sample densities ($N \in \{10, 20\}$ per skill).

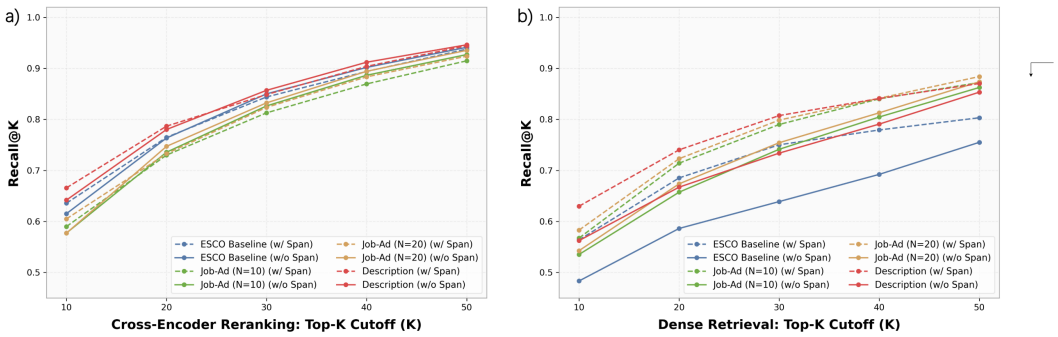


Figure 2: Effect of knowledge base expansion and span-based retrieval on Recall@K across candidate depths $K \in [10, 100]$ (w/o Span: without span-based retrieval).

From these results (Figure 2), we can draw several key observations. First, candidate reranking consistently outperforms dense retrieval across various Top-K cut-offs, showing the necessity and potential of reranking to effectively prune the candidate label pool while preserving high recall.

Second, regarding knowledge base construction, semantic expansion strategies prove beneficial primarily for dense retrieval. Specifically, LLM-synthesized skill descriptions emerge as the most effective expansion approach, while LLM-synthesized job advertisements also yield a considerable performance boost. This demonstrates the efficacy of LLM-generated data as a semantic expansion corpus for enriching knowledge base representations. However, these expansion strategies unexpectedly degrade reranking performance. This is likely because synthetic text introduces off-topic noise and dilutes high-density skill semantics. While dense retrieval benefits from broader semantic coverage, fine-grained rerankers are highly sensitive to such distractors, resulting in lower scores for true positive labels.

Third, span-based retrieval yields a significant performance boost for dense retrieval, whereas its impact on reranking remains marginal. This disparity can likely be attributed to the fact that dense retrieval relies heavily on localized, fine-grained text spans dense retrieval relies heavily on localized, fine-grained text segments to reduce background noise and capture specific skill mentions. Conversely, reranking models leverage full-context attention or generative reasoning. They naturally possess strong inherent capabilities to filter out irrelevant text and focus on key spans within the original context, rendering additional span-level segmentation largely redundant.

4.3 LLM-Based Classification

To evaluate LLM-based classification under controlled conditions, we constructed a simulated setup with a fixed candidate pool of 60 labels per query. Likewise, this pool comprises all gold labels, 30 hard negatives sampled from the most semantically similar false labels, and easy negatives randomly drawn from the broader skill ontology to fill the remaining capacity. For our experiments,

we selected `Gemini-2.5-Flash` and `DeepSeek-V4.1-Flash` as representative models, as both represent state-of-the-art architectures renowned for their strong reasoning capabilities and high efficiency in complex retrieval and reranking tasks.

Regarding the input candidate label pool size (Figure 3), a larger pool naturally encompasses a higher proportion of gold labels, theoretically providing greater potential for improved performance. However, the empirical results reveal a clear tipping point. Beyond an optimal size, performance declines as label overload occurs, diluting the model's attention mechanism with excess noise.

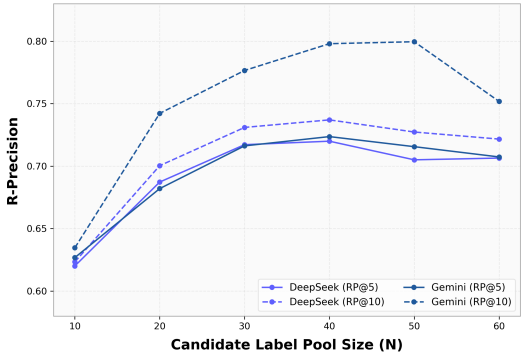


Figure 3: Sensitivity of LLM-based classification performance to input candidate label pool size ($K \in [10, 60]$) across Gemini and DeepSeek.

As for candidate label ordering (Table 1), the left columns depict the Recall and MRR metrics of the candidate pool fed into the LLM classifier, while the right columns present the final classification outcomes. Specifically, we evaluated six distinct ordering strategies, placing gold labels at the front, middle, and end of the candidate space, alongside random shuffling, dense retrieval, and reranking orderings. As shown by the classification results, label ordering exerts a substantial impact on final performance. Models perform markedly better when gold labels are situated near either the beginning or the end of the input sequence, which presents a pattern consistent with the "Lost-in-the-Middle" phenomenon. This underscores the necessity of utilizing effective reranking to actively position candidate gold labels toward the head (or tail) of the context window.

Table 1: Sensitivity of LLM-based classification performance to input candidate label ordering across Gemini and DeepSeek.

	Candidate Label Pool Distribution					**Gemini-2.5-Flash**		**DeepSeek-V4.1-Flash**	
Ordering	**R@10**	**R@20**	**R@30**	**R@60**	**MRR**	**RP@5**	**RP@10**	**RP@5**	**RP@10**
Gold-First	98.88	99.80	99.88	100.0	0.999	71.57	76.55	73.12	75.93
Gold-Last	0.000	0.010	0.020	100.0	0.017	70.10	74.21	71.50	75.56
Gold-Middle	0.000	0.060	77.65	100.0	0.035	68.59	73.15	68.34	70.26
Random	16.25	33.32	49.37	100.0	0.113	69.76	74.19	68.54	71.44
D-Retrieved	61.93	72.47	80.49	100.0	0.522	69.47	72.90	68.85	70.15
Reranked	68.93	84.72	93.31	100.0	0.539	70.73	75.18	70.64	72.16

To evaluate reasoning regimes (Table 2), we compare Chain-of-Thought (CoT) against classic zero-shot and few-shot prompting across native and non-native reasoning modes (see Figure A.2 in Appendix for details). Results indicate that CoT outperforms zero/few-shot prompting in non-native reasoning mode. However, under native reasoning mode, explicit reasoning prompts yield a significant adverse effect. We attribute this degradation to prompt competition and constraint redundancy, as explicit CoT instructions interfere with the model's internal, fine-tuned reasoning pathways, causing verbose output that disrupts exact label matching.

4.4 Main Results

Finally, we integrate all evaluated strategic components into the proposed Skill-RAG framework. Specifically, we first dense-retrieves the top-100 skill candidates from the full ESCO ontology,

Table 2: Impact of prompting strategies and explicit reasoning (thinking mode) on candidate reranking performance across models. (NR: Native Reasoning)

	Gemini-2.5-Flash				DeepSeek-V4.1-Flash			
	w/ NR		w/o NR		w/ NR		w/o NR	
Prompting Strategy	**RP@5**	**RP@10**	**RP@5**	**RP@10**	**RP@5**	**RP@10**	**RP@5**	**RP@10**
Zero-shot	70.40	79.63	63.60	66.12	66.88	70.04	66.34	69.53
Few-shot	**70.73**	**75.18**	65.04	67.81	70.64	72.16	68.64	71.33
Chain-of-Thought	65.60	73.02	69.11	75.54	64.95	66.33	**71.65**	**79.24**

reranks them to the top-60 using a cross-encoder, and finally prompts the LLM while incorporating all aforementioned optimal strategies (i.e. Knowledge Base Semantic Expansion, Span-Based Retrieval, Chain-of-Thought) to maximize performance. Following established literature, we report evaluation results separately across the three SkillSpan benchmark subsets: HOUSE, TECH, and TECHWOLF. Compared with previous studies utilizing deep learning-based approaches (Decorte et al., 2022) and LLM-based architectures (Decorte et al., 2023; Clavié & Soulié, 2023; D'Oosterlinck et al., 2024b), our top-performing configuration sets a new state-of-the-art across market benchmarks (Table 3). In summary, Skill-RAG establishes a computationally efficient and accurate pipeline for extreme multilabel skill classification.

Table 3: Performance of Skill-RAG compared with different previous classification regimes.

	SkillSpan								
Sub-Dataset	**HOUSE**			**TECH**			**TECHWOLF**		
Metric	**MRR**	**RP@5**	**RP@10**	**MRR**	**RP@5**	**RP@10**	**MRR**	**RP@5**	**RP@10**
Previous Study									
Decorte et al. (2022)	0.296	30.57	–	0.325	32.12	–	0.284	29.27	–
Decorte et al. (2023)	0.428	45.74	–	0.529	54.62	–	0.526	54.57	–
Clavié & Soulié (2023)	0.507	56.67	60.09	0.537	61.50	68.94	–	–	–
D'Oosterlinck et al. (2024b)	–	56.50	65.76	–	59.61	70.23	–	57.04	65.17
SKILL-RAG									
Skill-RAG (Base)	0.597	60.94	63.97	0.603	60.57	63.09	0.605	57.59	59.26
+ Span-Retrieval & KB Expansion	0.608	63.59	68.79	0.637	65.61	69.01	0.618	61.38	67.36
+ Chain-of-Thought	0.600	62.43	70.76	0.638	65.08	71.88	0.598	62.89	67.76
+ All	0.637	65.14	71.76	0.637	67.49	72.27	0.613	64.01	68.58

5 Conclusion, Limitations, and Future Work

In this work, we present Skill-RAG, a computationally efficient and accurate pipeline tailored for extreme multilabel skill classification. By systematically exploring knowledge base semantic expansion, span-based retrieval, input candidate pool sizing, candidate label ordering, and LLM reasoning regimes, our framework sets a new state-of-the-art across standard market benchmarks while significantly alleviating reliance on labor-intensive manual annotations. Notably, our findings reveal critical insights into the adverse effects of redundant CoT prompting on native reasoning architectures, offering concrete guidance for future RAG and LLM classification designs.

Despite these contributions, our study has certain limitations that open avenues for future work. First, our label ordering analyses focus primarily on static candidate pool sizes (e.g., $K = 60$), leaving dynamic, adaptive pool allocation under extreme scale to be further explored. Second, while Skill-RAG has been primarily evaluated on ESCO and SkillSpan due to dataset accessibility, its effectiveness across broader skill classification scenarios still requires further exploration. Future research will focus on developing adaptive candidate-ordering strategies, exploring parameter-efficient fine-tuning for native reasoning models, and extending Skill-RAG to multilingual and streaming skill classification scenarios.

AI USE STATEMENT

We used generative AI tools only for language editing and limited programming assistance, including improving grammar and readability, debugging code, and checking code formatting. We did not use generative AI to develop research questions or methods, conduct literature reviews, analyze or interpret data and results, create figures or tables, or generate references or other research materials. All AI-assisted text and code were reviewed by the authors. Code suggestions were checked and tested before use. The authors take full responsibility for the final manuscript and all reported results.

ETHICS STATEMENT

This work does not involve human participants, interventions, or the collection of personal data. We used only data that were publicly available or otherwise accessed and processed in accordance with their applicable terms of use. And we did not collect, generate, or release sensitive personal information. We did not use the proposed method to make decisions about individuals or groups. The method may be affected by limitations or biases in the data and models used. We therefore encourage users to evaluate its performance and potential bias carefully before applying it in high-stakes settings. The authors take responsibility for the integrity of the study, the reported results, and compliance with applicable research and data-use requirements.

REPRODUCIBILITY STATEMENT

We provide all necessary details to ensure the reproducibility of our work across the main text and the Appendix, including complete task specifications, dataset characteristics, methodological implementations, and evaluation procedures. Furthermore, we have uploaded our anonymized codebase, prompt templates, and data-processing scripts as supplementary material, which are also available at `https://anonymous.4open.science/r/S_C-63FF`. All resources are released in full compliance with the licensing terms and conditions of the underlying datasets.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chenxin An, Jiahuan Zhang, Ming Zhong, Linzheng Li, Shansan Gong, Yuan Luo, Jing Xu, and Lingpeng Kong. Why does the effective context length of llms fall short? *arXiv preprint arXiv:2410.18745*, 2024.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of LLMs fall short? In *International Conference on Learning Representations*, 2025.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12119–12149, 2025. doi: 10.18653/v1/2025.naacl-long.605.

Akash Bhola, Kishaloy Halder, Amrit Prasad, and Min-Yen Kan. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5832–5842, 2020. doi: 10.18653/v1/2020.coling-main.513.

Peter Q. Blair, Papia Debroy, and Justin Heck. Skills, degrees, and labor market inequality. *Economics of Education Review*, 108:102704, 2025.

Francesco Busolin, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Salvatore Trani, and Alberto Veneri. Efficient re-ranking with cross-encoders via early exit. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. doi: 10.1145/3726302.3729962.

Federica Calanca, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. *EPJ Data Science*, 8(1):13, 2019. doi: 10.1140/epjds/s13688-019-0190-z.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pp. 78–87, 2019. doi: 10.18653/v1/W19-2209.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3163–3171, 2020. doi: 10.1145/3394486.3403368.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Benjamin Clavié and Guillaume Soulié. Large language models as batteries-included zero-shot ESCO skills matchers. *arXiv preprint arXiv:2307.03539*, 2023.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024. doi: 10.1145/3626772.3657834.

Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. Design of negative sampling strategies for distantly supervised skill extraction. *arXiv preprint arXiv:2209.05987*, 2022.

Jens-Joris Decorte, Severine Verlinden, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. Extreme multi-label skill extraction training using large language models. *arXiv preprint arXiv:2307.10778*, 2023.

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.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*, 2024a.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*, 2024b.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024. doi: 10.1145/3637528.3671470.

Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu, Fuzhen Zhuang, and Hui Xiong. RecruitPro: A pretrained language model with skill-aware prompt learning for intelligent recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3991–4002, 2023. doi: 10.1145/3580305.3599894.

Joshua Fields, Kevin Chovanec, and Praveen Madiraju. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531, 2024. doi: 10.1109/ACCESS.2024.3349952.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Nidhi Goyal, Jushaan Kalra, Charu Sharma, Raghava Mutharaju, Niharika Sachdeva, and Ponnurangam Kumaraguru. JobXMLC: EXtreme multi-label classification of job skills with graph neural networks. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2181–2191, 2023. doi: 10.18653/v1/2023.findings-eacl.163.

Akshay Gugnani and Hemant Misra. Implicit skills extraction using document embedding and its use in job recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13286–13293, 2020. doi: 10.1609/aaai.v34i08.7038.

Xinyu Guo, Yazhou Zhang, and Jing Qin. Textreasoningbench: Does reasoning really improve text classification in large language models? *ArXiv.org*, March 2026. URL https://arxiv.org/abs/2603.19558.

Amirhossein Herandi, Yitao Li, Zhanlin Liu, Ximin Hu, and Xiao Cai. Skill-LLM: Repurposing general-purpose LLMs for skill extraction. *arXiv preprint arXiv:2410.12052*, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024a.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024b. doi: 10.18653/v1/2024.findings-acl.890.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8153–8161, 2021. doi: 10.18653/v1/2021.emnlp-main.643.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. LightXML: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7987–7994, 2021. doi: 10.1609/aaai.v35i9.16974.

Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. Skills prediction based on multi-label resume classification using CNN with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087, 2021. doi: 10.1007/s00521-020-05302-x.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Imane Khaouja, Ismail Kassou, and Mounir Ghogho. A survey on skill identification from online job ads. *IEEE Access*, 9:118134–118153, 2021. doi: 10.1109/ACCESS.2021.3106120.

Siddhant Kharbanda, Atmadeep Banerjee, Erik Schultheis, and Rohit Babbar. CascadeXML: Rethinking transformers for end-to-end multi-resolution training in extreme multi-label classification. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, 2024. doi: 10.18653/v1/2024.acl-long.818.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Nan Li, Bo Kang, and Tijl De Bie. SkillGPT: a RESTful API service for skill extraction and standardization using a large language model. *arXiv preprint arXiv:2304.11060*, 2023.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context LLMs struggle with long in-context learning. *Transactions on Machine Learning Research*, 2025.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 115–124, 2017. doi: 10.1145/3077136.3080834.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2472–2495, 2025. doi: 10.18653/v1/2025.emnlp-main.126.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. JobSkape: A framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pp. 43–58, 2024. doi: 10.18653/v1/2024.nlp4hr-1.4.

Lorenzo Malandri, Fabio Mercorio, and Antonio Serino. SkiLLMo: Normalized ESCO skill extraction through transformer models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pp. 1969–1978, 2025. doi: 10.1145/3672608.3707960.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. Label semantic aware pre-training for few-shot text classification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8318–8334, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.570. URL https://aclanthology.org/2022.acl-long.570/.

Khanh Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. Rethinking skill extraction in the job market domain using large language models. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pp. 27–42, 2024. doi: 10.18653/v1/2024.nlp4hr-1.3.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718, 2020. doi: 10.18653/v1/2020.findings-emnlp.63.

Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3980–3990, 2019. doi: 10.18653/v1/D19-1410.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022. doi: 10.18653/v1/2022.naacl-main.191.

Duxin Shang, Yufeng Duan, Ping Bai, and Jiahong Xie. Exploring retrieval-augmented generation for multi-label discipline classification of academic short texts. *Scientometrics*, 131(4):2373–2399, 2026. doi: 10.1007/s11192-025-05472-2.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, 2025.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, 2023a. doi: 10.18653/v1/2023.emnlp-main.923.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8990–9005, 2023b. doi: 10.18653/v1/2023.findings-emnlp.603.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Niek Vermeer, Vera Provatorova, David Graus, Thilina Rajapakse, and Sepideh Mesbah. Using robbert and extreme multi-label classification to extract implicit and explicit skills from dutch job descriptions. *Compjobs' 22: Computational Jobs Marketplace*, 1(1):2–6, 2022.

Huimu Wang, Yiming Qiu, Xingzhi Yao, Guangtao Nie, Zuxu Chen, Zhenlin He, Songlin Wang, Guoyu Tang, Sulong Xu, Jingwei Zhuo, et al. Bridging the gap: Generative retrieval via query-to-multi-span framework for effective e-commerce search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4960–4964, 2026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Zihan Wang, Yuxuan Lin, Jun Shen, and Xiaofeng Zhu. A survey of large language models for text classification: What, why, when, where, and how. *Authorea Preprints*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. In *International Conference on Learning Representations*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. In *International Conference on Learning Representations*, 2025.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit S. Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7267–7280, 2021.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. SkillSpan: Hard and soft skill extraction from English job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4962–4984, 2022. doi: 10.18653/v1/2022.naacl-main.366.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024. doi: 10.18653/v1/2024.acl-long.814.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *International Conference on Learning Representations*, 2023.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of LLMs a mirage? a data distribution lens. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15231–15261, 2026.

Meng Zhao, Faizan Javed, Ferosh Jacob, and Matt McNair. SKILL: A system for skill identification and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 4012–4017, 2015. doi: 10.1609/aaai.v29i2.19064.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4), 2024. doi: 10.1145/3637870.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.

A Appendix

A.1 Prompt Templates

Instruction (Skill Description)

[Task Instructions]
- Given: A skill label name.
- Your goals: Generate a comprehensive skill description that captures its core semantic features, including its formal definition, typical job responsibilities, associated tools or technologies, and relevant context.

[Output Format]
Return only the generated skill description.

----- k demonstrations -----

Instruction (Job-Ad Sample)

[Task Instructions]
- Given: A skill label name.
- Your goals: Generate ***N*** realistic job ad sample sentences that effectively capture the core semantic features, key responsibilities, practical application contexts, and typical co-occurring technical requirements of the skill.

[Output Format]
Return matched job ad samples as a list of strings in this exact format: ['Job_Ad_Sample_1'; 'Job_Ad_Sample_2'; ...].

----- k demonstrations -----

Figure A.1: Prompt Template for Knowledge Base Expansion via LLM Corpus Synthesis

Instruction (Zero-Shot Style)

[Task Instructions]
- Given:1. A job ad text. 2. A 'Candidate Skill Label Space' with the form: Skill <ID>: <short title>: <natural language definition>.
- Your goals: Identify all skills from the given 'Candidate Skill Label Space' that are mentioned in the job ad.

[Output Format]
Return matched skills as a list of IDs in this exact format: ['skill_1_ID';'skill_2_ID'; ...]. If no matches exist, return: 'NaN'.

Instruction (Few-Shot Style)

[Task Instructions]
- Given:1. A job ad text. 2. A 'Candidate Skill Label Space' with the form: Skill <ID>: <short title>: <natural language definition>.
- Your goals: Identify all skills from the given 'Candidate Skill Label Space' that are mentioned in the job ad.

[Output Format]
Return matched skills as a list of IDs in this exact format: ['skill_1_ID';'skill_2_ID'; ...]. If no matches exist, return: 'NaN'.

----- k demonstrations -----

Instruction (Chain-of-Thought)

[Task Instructions]
Given:1. A job ad text.2. A 'Candidate Skill Label Space', which is a list of skill entries. Each entry has the form: Skill <ID>: <short title>: <natural language definition>.
Your goals: Identify all skills from the given 'Candidate Skill Label Space' that are mentioned in the job ad.

[Reasoning Procedure]
1. Extract skill-related text spans - Scan the entire job description.- Identify all phrases/sentences explicitly describing skills, abilities, responsibilities.- List them as: 'Span 1: "quote text"'; 'Span 2: "quote text"', etc.
2. Match each span to skills - For each span, Based on its original background context, understand the skill it represents. - In Skill Label Space, search for skills where name OR definition closely matches the span's meaning, while considering the full context of the job description.- List matches: 'Skill ID: reason why it matches this span'.
3. Compile final list - Remove duplicates across all spans.- Only include skills with clear evidence.

[Output Format]
The output should consist of the Model's Reasoning text and Result.
For the Reasoning text: Please first output the skill-related text spans extracted, then the matching process
For the Result:
- If at least one skill is relevant: '['ID_1';'ID_2'; ...]' (for example: ' ['13295';'6142';'1171'] ').
- If no skills are relevant: ['NaN']

----- 2 demonstrations -----

Figure A.2: Prompt Template for LLM-Based Classification