Enhancing Job Matching: Occupation, Skill and Qualification Linking with the ESCO and EQF taxonomies

Anonymous ACL submission

Abstract

This study investigates the optimal utilization of Large Language Models (LLMs) for linking job vacancy texts to the ESCO taxonomy and the EQF classification. We demonstrate that an entity-linking methodology significantly outperforms traditional sentence similarity approaches, and we release our entity linker to facilitate further research. To advance beyond skill extraction, we introduce two novel datasets for evaluating occupation and qualification extraction. Furthermore, we explore optimal embedding strategies for ESCO nodes in a retrieval setting, revealing which combination of fields is the most effective for occupations and which works best for skills. Finally, we achieve state-of-the-art results on an established dataset for job entity extraction.

1 Introduction

Recent developments in deep learning have spurred significant advancements in the job domain. This emerging field emphasizes the skill extraction paradigm, wherein deep neural networks are employed to extract skill-related information from plain-text job vacancies (Senger et al., 2024). However, these texts also contain various other types of information—such as occupations and qualifications—that warrant further attention. We argue that robust models should not only identify these additional entities but also, where feasible, link them to an appropriate knowledge base.

Linking job descriptions to established taxonomies—such as the European Skills, Competences, Qualifications and Occupations (ESCO) (le Vrang et al., 2014) or the International Standard Classification of Occupations (ISCO)—remains a pivotal challenge. Early Large Language Models (LLMs) have demonstrated effectiveness in extracting robust semantic representations from unstructured text, as shown by Devlin et al. (2018). Building on this foundation, sentence embedding techniques introduced in SBERT (Reimers and Gurevych, 2019) have further enhanced the efficiency of text classification and semantic similarity tasks, which are critical for mapping job descriptions to standardized occupational frameworks.

In this work, we address the following research question: What is the optimal way to employ LLMs for linking job descriptions to established taxonomies?

We select the ESCO taxonomy as our use case and aim to match job vacancy texts to its nodes. This constitutes a text classification problem, for which we investigate two possible approaches.

(Methodology 1) We approach the task as a *sentence linking* (SL) problem, feeding complete job descriptions into the models and expecting a list of ESCO nodes as outputs. This approach is often labeled as extreme multi-label classification (Decorte et al., 2023; D'Oosterlinck et al., 2024).

(Methodology 2) We introduce an intermediate step where the models first perform entity recognition (ER) (Li et al., 2022), thereby framing the task under the *entity linking* (EL) paradigm (Sevgili et al., 2022).

We explore both methodologies and present a comparative analysis using transformer-based neural networks as our foundation.

Previous studies have primarily concentrated on skill extraction from job vacancies, often overlooking other job-related entities. This limitation is likely due to the inherent complexity of the broader task. Defining what precisely constitutes a "skill" is itself challenging, introducing ambiguities into the training data.

Some prior work has adopted Methodology 1, applying plain sentence similarity strategies focused solely on skills. For example, Khaouja et al. (2021) compare using sent2vec trained on Wikipedia sentences with SBERT, which is trained on large collections of paraphrased sentences to generate embeddings. Similarly, Zhang et al. (2022b) employ

language models to align n-grams extracted from job postings with the ESCO taxonomy. Furthermore, Decorte et al. (2023) and Clavié and Soulié (2023) utilize a synthetic skills training set to directly link sentences with skills, employing LLM-based re-rankers. In the work of Gnehm et al. 2022, skill extraction is conducted directly by leveraging context-aware embeddings and the SBERT model, in a manner similar to Zhang et al. (2022b). Moreover, their approach contextualizes skill domains within specific spans and ontology terms, utilizing ESCO's hierarchical structure.

In contrast, Methodology 2 has not received as much academic attention. An EL paper focused on the job domain was published by Zhang et al. (2024), in which the authors train two widely used models for this task: BLINK (Wu et al., 2019) and GENRE (De Cao et al., 2020). They assess the effectiveness of skill extraction using this methodology with synthetic training data provided by Decorte et al. (2023), achieving moderate yet promising results. The authors emphasize the need for a more comprehensive dataset for evaluation. In our work, we use the same evaluation set, introduced by Decorte et al. (2022), for the skills component of our study.

SkillGPT (Li et al., 2023) represents the first tool to employ a large language model (LLM) for the matching task. It transforms ESCO entries into structured documents, which the language model subsequently vectorizes. The input job text is then condensed into a summary, and the embedding of this summary is used to retrieve the most relevant ESCO entries. SkillGPT's architecture resembles an EL pipeline, as it follows a two-step process. Although it incorporates both skill and occupation entities, the authors unfortunately do not provide an analytical evaluation.

Given the substantial progress in recent years, we aim to advance the research field by proposing an evaluation framework for skill, occupation, and qualification extraction concerning ESCO and EQF.

Data scarcity remains a significant challenge in the job domain when applying machine learning algorithms. To address this issue, we introduce three novel datasets: one for evaluating occupation linking with the ESCO taxonomy, another for qualification linking to the European Qualifications Framework (EQF), and a third for assessing occupation title similarity. A detailed description of all datasets used in this study is provided in Section 2.

When considering a Retrieval-Augmented Generation (RAG) architecture (Gao et al., 2023), both of the aforementioned methodologies can serve as the retrieval component of the system. Given the growing popularity of RAG and in-context learning (D'Oosterlinck et al., 2024; Kavas et al., 2025), it is essential to examine the respective strengths and weaknesses of these approaches. This constitutes the primary motivation behind our research.

Additionally, we conducted extensive experiments on entity extraction using a well-established benchmark: the dataset introduced by Green et al. (2022). We achieve state-of-the-art results on this benchmark, which are reported in Section 4.

Overall, the EL approach produced the most effective results, as it facilitates a more precise information flow—embedding only the most relevant textual segments via text embedding models. Further analysis can be found in Section 5.

As a byproduct of this comparison, we also investigated optimal strategies for embedding ESCO nodes in a retrieval context. Each node consists of multiple data fields, which opens the door to diverse embedding techniques. We present our findings in Section 3.

Finally, in Section 6, we explore various strategies for leveraging the latest generation of LLMs to support our task.

2 Datasets

In this section, we present the datasets used throughout our work. These are categorized into three groups: reference sets, evaluation sets, and training sets. Detailed data statistics are provided in Appendix B.

2.1 Reference Sets

ESCO The central aim of this study is to classify arbitrary English-language job vacancy texts using the ESCO taxonomy.

We utilize version 1.1.1 of ESCO, which contains 3,007 Occupations and 13,896 Skills. Both the Skill and Occupation frameworks are organized as taxonomies (Poli et al., 2010)—that is, they follow subclass relationships—where each Skill may have multiple parent categories. In this work, we focus exclusively on discrete entities within ESCO and disregard hierarchical relationships between broader concepts or links between Occupations and Skills. We leave this aspect for future exploration.

EQF ESCO defines a qualification as the official outcome of an assessment by a competent body that verifies an individual's learning achievements against established standards (ESCO, 2024). The qualification data available in Europass are sourced from national databases representing the frameworks of EQF member countries. Europass offers a consolidated repository of current, high-quality data on qualifications, national frameworks, and educational trajectories across Europe (Europass, 2024). We extract relevant information on EQF levels from the official European Union comparison portal.[1] Only English-language content is retained. This results in a dataset of 814 entries, each consisting of a qualification string, the issuing country, and the corresponding EQF level (Table 8).

2.2 Evaluation Sets

Ethiopian Dataset To evaluate occupational classification, we employ a dataset comprising job descriptions annotated with corresponding ESCO occupation codes.

The vacancy data were collected from both online and offline sources in Ethiopia. Offline sources include physical job boards, public postings, and government gazettes across major cities. Online sources involve local job portals, an Ethiopian enterprise platform, and digital media managed by employers. Data are gathered either directly via the Ethiopian platform or through web scraping. In addition, printed job advertisements are photographed at the Ethiopian employment center for digital processing.

All collected data are reviewed and annotated by trained personnel using proprietary tools. Staff members receive specialized training on ESCO, ISCO, and O*NET classification systems, covering taxonomy structure, application rationale, and practical annotation exercises.

We compile real-world evaluation sets (Table 9) for each entity type relevant to our models.

Occupations We use a subset of the Ethiopian Jobs dataset containing 542 annotated entries (Table 9), each comprising a job title, a job description, and the relevant ESCO occupation code. This subset is constructed ensuring diversity across multiple job sectors.

Skills For skill evaluation, we utilize the dataset introduced by Decorte et al. (2022), which includes the HOUSE and TECH extensions of the SkillSpan dataset (Zhang et al., 2022a). These datasets feature test and development sets with SkillSpan entities mapped to the ESCO model.

Qualifications We extend the Green Benchmark Qualifications dataset by mapping each entry to the appropriate EQF level.

Two native Greek-speaking annotators (one male and one female) performed the annotation process. The resulting inter-annotator agreement, measured using Cohen's Kappa (Fleiss and Cohen, 1973), was 0.45—indicating moderate agreement. Qualifications that did not align with any EQF level were labeled as unknown (UNK). A common example is the "driving license," which is not associated with any EQF level under ESCO.

To improve consistency, we resolve disagreements as follows: when both annotators select valid but differing EQF levels, we assign the lower level. If one annotator selects UNK while the other provides a valid EQF level, we consult `Gemini 1.5 Pro` as an adjudicator. The model is prompted to choose between the two annotations, and its decision is included in the final dataset. Details of the prompts used are provided in Appendix C.

2.3 Training Sets

To support Methodology 2, we train entity extraction models using the benchmark dataset introduced by Green et al. (2022) (Table 6).

Title Similarity Dataset To enhance occupational classification performance, we further finetune two sentence transformers using a derivative dataset from the Ethiopian Dataset. We construct this dataset (Table 7) by aligning job titles with the preferred and alternative labels specified in the ESCO occupation taxonomy.

3 Methodology #1 : Sentence Linking

Let D be the Document space and a Sentence Transformer $ST : D \rightarrow \mathbb{R}_n$ be an embedding function to an arbitrary Euclidean metric space. Also, let $O = \{o_1, o_2, ..., o_{3007}\}$, $S = \{s_1, s_2, ..., s_{13896}\}$ and $Q = \{q_1, q_2, ..., q_{814}\}$ be the reference sets described in section 2. Our goal is to retrieve entities from these sets so we embed O, S and Q using ST and cache them in separate vector databases. We define a *query* which is a plain-text sentence, annotated with entities from the reference sets.

We consider different possible ways of embedding the ESCO occupations and skills nodes and of comparing the embedding to the query, to find

[1] https://europass.europa.eu/en/compare-qualifications

the one that maximizes precision. With respect to the Skills and Qualifications Evaluation Sets, we remove the UNK labels and link each sentence only one time.

To improve Occupational matching, provided the title similarity dataset described in section 2, we fine-tune the `all-mpnet-base-v2` sentence transformer. The model was trained on minimizing the Multiple Negatives Loss (Henderson et al., 2017), using default hyperparameters.

The relevant textual fields for each ESCO node are: *preferred label*, *description*, *secondary labels*, i.e., alternative titles presented as a newline-separated list.

To evaluate various embedding strategies, we consider the following configurations: (1) Single embedding: *preferred label* (2) Single embedding: *description* (3) Single embedding: concatenation of *preferred label* and *description* (4) Single embedding: concatenation of all fields into a single string (5) Multiple embeddings: one per field (*preferred label*, *description*, and combined *secondary labels*) (6) Multiple embeddings: one for *preferred label*, one for *description*, and individual embeddings for each *secondary label*.

In the multiple embedding setup, retrieval is based on the highest cosine similarity between any field-level embedding and the query. The top-k nodes are selected based on these maximum similarities, ensuring duplicate entries are removed.

The results of this experiment are summarized in Table 10, with evaluation based on Accuracy@1, in line with prior work (Zhang et al., 2024; Zaporojets et al., 2022).

For occupations, we find that multi-field embeddings (strategy 5) improve performance—provided that the inclusion of secondary labels does not introduce excessive noise or redundancy (due to overlapping labels across nodes). In contrast, for skills, injecting too much information via multiple fields degrades performance. The optimal strategy is to embed the concatenation of all fields (strategy 4). Notably, embeddings based solely on the preferred label offer nearly comparable accuracy while reducing computational overhead, which is especially relevant given the large number of ESCO skills.

Fine-tuning on occupation-specific data significantly improves accuracy for the Occupations task, without negatively impacting performance on Qualifications. However, for Skills, we observe a drop in performance after fine-tuning, suggestive of catastrophic forgetting.

Next, we investigate whether an ER approach outperforms full-sentence embedding. As a preliminary analysis, we rerun the previous experiment using job titles (Title Linking) as queries, and compare their embedding-based retrieval performance against the earlier configurations. Results are presented in Table 11.

We observe a substantial improvement in performance, with an approximate 15% increase in accuracy. These findings reinforce our motivation to develop a dedicated EL model to surpass our current SL baseline.

4 Methodology #2: Entity Linking

Given an input text document $D = \{w_1, ..., w_r\}$ and a list of entity mentions (n-grams corresponding to entities) $M_D = \{m_1, ..., m_n\}$, the output of an EL model is a list of mention-entity pairs $\{(m_i, e_i)\}, i \in [1, n]$. Each entity e_i is an element in a set E of all possible entities in a knowledge base (e.g. WikiData, DBpedia, ESCO).

Most EL-related works hypothesize that the mentions are explicitly given in the training and test datasets. Inspired by Sevgili et al. (2022), we distinguish the mention detection and entity disambiguation steps and assume that the mention boundaries are missing from the evaluation procedure. Consequently, as shown in Figure 1, this model consists of two discrete modules.

Entity Recognition Module Formally, the ER task according to Zhang et al. (2022b) is defined as follows. Let d be a subset of sentences (sequences of tokens) from a job posting D. Let $X_d^i = \{x_1, x_2, ..., x_T\}$ be the i^{th} sequence of input tokens and $Y_d^i = \{y_1, y_2, ..., y_T\}$ be the target sequence of BIO labels (e.g., "B-Skill", "I-Occupation", "O") corresponding tp this input sequence. The goal is to use D to train a sequence labeling algorithm $h : X \rightarrow Y$ to accurately predict entity spans by assigning an output label y_t to each token x_t.

We perform the ER task by training BERT-based models for token classification. We experimented with language models of various sizes and pre-training schemes. Namely, we used `BERT` with both its base and large variances on the cased version. Also, we experimented with two domain-adapted models, `JobBERT` (Zhang et al., 2022a) and `ESCOXLM-R` (Zhang et al., 2023) to test whether domain adaptation generalizes in our holistic overview of job postings text analysis. Both

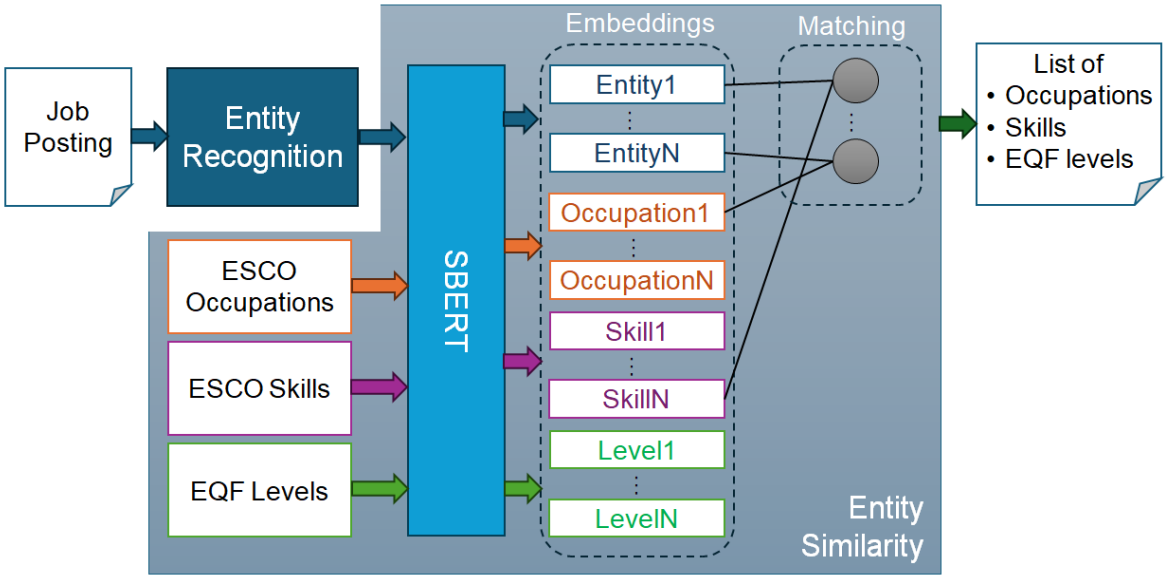

Figure 1: Entity Linking Job Posting Analysis Framework

RoBERTa$_{base}$ and RoBERTa$_{large}$ (Liu et al., 2019) were fine-tuned on our task, as well as the first version of Microsoft's DeBERTa$_{base}$(He et al., 2020) model.

Based on previous work (Zhang et al., 2022a; Souza et al., 2019; Jensen et al., 2021), we experimented adding a conditional random field (Lafferty et al., 2001) decoder on top of transformer language models for improved accuracy.

Entity Similarity Module Let (m_i, e_i) be a tuple of an extracted mention by the entity extractor, where $m_i \in \mathcal{P}(D)$, $e_i \in E$. D is the Document space, $\mathcal{P}(D)$ it's power set and E is the set of entity categories. Similar to section 3, we represent Occupations O, Skills S, and Qualifications Q using a Sentence Transformer (ST) to generate the corresponding embedding vectors in $\mathbb{R}^n$ space. Note that $E = O \cup S \cup Q$.

Given a job posting $d = \{w_1, ..., w_n\}$, we apply the NLTK (Bird et al., 2009) package to tokenize the document into chunks, $X_k = \{x_1, ..., x_k\}$. Each X_k is passed through the ER function h to generate the BIO labels $h(X_k) = Y_k = \{y_1, ...y_k\}$. From these we can obtain the mentions m_i and apply post-processing steps to improve performance.

These steps include: (1) removing special tokens (e.g. [SEP], [CLS], <s>, etc), (2) correcting common sequence errors such as converting the sequence (..., "B-", "O", "I-",...) to (..., "B-", "I-", "I-",...), and (3) ignoring single "I-" tags appearing at the end of a sentence.

For each mention m_i the Sentence Transformer is used to generate the embedding vector $\mathbf{V} = ST(m_i) \in \mathbb{R}^n$.

We then proceed to compute the cosine similarity of $\mathbf{V}$ against $o_j \in O$, $s_i \in S$ and $q_k \in Q$, depending on the category indicated by the ER module. Finally, we retrieve ranked lists of the top-k ESCO Occupations, ESCO Skills or EQF Qualification entities based on the above metric.

We experiment with two sentence transformers: `all-MiniLM-L6-v2` and `all-mpnet-base-v2`. The `all-MiniLM-L6-v2` model is further fine-tuned on the title similarity set, similar to `all-mpnet-base-v2` as described in Section 3. All four resulting models are evaluated as candidates for our sentence transformer function ST.

4.1 Evaluation

The ER training was assessed using standard span F1 strict metric (Li et al., 2022; Nakayama, 2018), where true positives are considered if the exact entity span is predicted.

The entity similarity evaluation can be categorized into *in-KB Evaluation* when all the entities in the evaluation set are from the same knowledge base, and *out-of-KB Evaluation* when Unknown labels correspond to entities in the text.

For out-of-KB Evaluation, we developed an algorithm using the whole system to evaluate the similarity module. Specifically, based on the extracted

entities on a given evaluation set, we check whether an overlap exists with the ground truth entity using the Jaccard Similarity (Jaccard, 1912). The ground truth span that maximizes the Jaccard Similarity with the extracted entity is then attributed to the top-k retrieved entities from the reference sets. If no overlap exists, the system returns the Unknown (UNK) label. Furthermore, the system returns UNK, if the retrieved item with the highest cosine similarity does not exceed a predetermined limit. After multiple experiments, we set this limit to 0.7 for the Skills and 0.8 for the Qualifications. When we perform the in-KB evaluation, the limit is set to 0 for all entities.

As a metric, we use the Mean Average Precision (MAP) evaluated at the first and fifth positions of the recommendations. MAP offers a single-figure measure of quality across different levels of recall. It is particularly noted for its excellent discrimination and stability. Given a set of input queries C, we calculate:

$$MAP@k(C) = \frac{\sum_{c=1}^{|C|} AvgP@k(c)}{|C|}$$

Average Precision at k is computed using the formula:

$$AvgP@k = \frac{\sum_{i=1}^{k} P(i) \times rel(i)}{\text{number of relevant documents}}$$

where i is the rank in the sequence of retrieved documents k is the number of retrieved documents, $P(i)$ is the precision at cut-off k in the list, $rel(i)$ is an indicator function equaling 1 if the item at rank i is a relevant document, zero otherwise. This metric is chosen since MAP does not penalize the suggestions if few relevant items exist. For consistency in subsequent comparisons, we define $Accuracy@1 := MAP@1$.

4.2 Experiments

Entity Recognition Our system's foundation is the ER module, which acts as the mention detector in the EL framework. Similar to traditional EL models (Sevgili et al., 2022; Stern et al., 2012), the ER errors propagate to entity disambiguation.

All training was conducted using V100 GPUs provided by a trusted source[2]. For the ER training, we performed a comprehensive grid search over hyperparameters for all encoder models. In all experiments, we utilized the test set from Green et al. (2022) for both validation and evaluation. The best-performing configurations were selected as outlined in Appendix A.

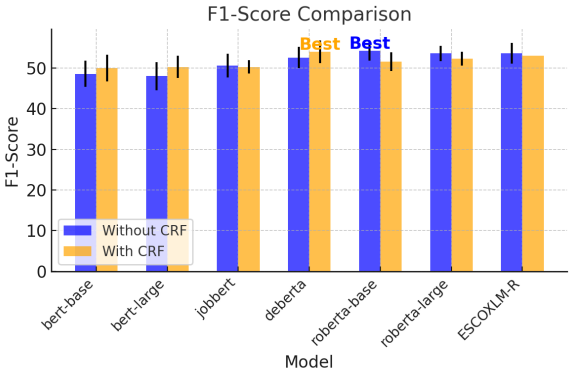

Figure 2: F1-Score results on the Green Benchmark test set. The results show the mean and standard deviation over three random seeds.

Figure 2 presents the results of the entity extraction experiments conducted on the Green Benchmark dataset. Each model was trained using three different random seeds to ensure robustness. The optimal model for this dataset is $\texttt{RoBERTa}_{\text{base}}$, which achieves a strict F1-score of **54.3**$\pm$2.6.

Overall, we observe that the addition of a CRF decoder enhances the performance of both BERT and DeBERTa models, but does not yield improvements for RoBERTa.

Our best-performing model, $\texttt{RoBERTa}_{\text{base}}$, establishes a new state-of-the-art on this benchmark. In our experimental setup, the previously reported state-of-the-art model, ESCOXLM-R (Zhang et al., 2023), achieves an F1-score of 53.6 ± 2.5 F1-score.

Entity Similarity We represent ESCO nodes based on the findings in Section 3: for Occupations, we adopt the multiple embeddings—one per field strategy, while for Skills, we use single embedding: preferred label. The evaluation sets are assessed using fourteen fine-tuned entity extractors in combination with four sentence transformers. We report the best-performing models for each entity type in Table 1, using the MAP metric.

While occupational fine-tuning significantly enhances performance on the Occupations dataset, it leads to a substantial drop in MAP when evaluated on Skills and Qualifications. This suggests the presence of catastrophic forgetting and indicates a need for a more diverse and representative training set to mitigate this effect.

[2]To be disclosed following the review process

Entity Type	Entity Model	Similarity Model	MAP@1	MAP@5
out-of-KB				
Skills	roberta-base	all-mpnet-base-v2	0.497	0.494
EQF	bert-large-cased	all-mpnet-base-v2	0.640	0.630
in-KB				
Occupations	roberta-large+CRF	all-mpnet-base-v2-FT	0.489	0.375
Skills	roberta-base+CRF	all-MiniLM-L6-v2	0.326	0.387
EQF	bert-large-cased	all-MiniLM-L6-v2	0.350	0.203

Table 1: **Entity Linking Results** With FT (fine-tuned) we note the models that were fine-tuned on the Ethiopian training set. *in-KB Evaluation* refers to the absence of unknown (UNK) labels in evaluation sets.

In-KB Evaluation When assuming that all evaluation entities belong to the known entity set E, we observe a slight decrease in MAP. This suggests that UNK labels, which are prevalent in both the Skill and Qualification datasets, impact the evaluation outcome. Notably, several prior studies (Clavié and Soulié, 2023; Decorte et al., 2023; D'Oosterlinck et al., 2024) evaluating the Skill dataset do not explicitly describe their handling of UNK labels.

Out-of-KB Evaluation In this setting, our results regarding the Skills evaluation set can be directly compared to Zhang et al. (2024), who report 23.55% Accuracy@1 for their best model. Our best approach achieves 49.7%, though this result may be biased due to the treatment of UNK labels. Unlike Zhang et al. (2024), who evaluate only entity similarity and disambiguation, our system performs both entity extraction and disambiguation. We advocate for a more comprehensive evaluation methodology that jointly assesses both tasks while preserving UNK labels (akin to Kolitsas et al. (2018)), as this more accurately reflects real-world conditions for entity linking applications.

5 Methodologies comparison

We must denote that the evaluation on sentence linking is done on the sentence level, while the EL is done on the entity level, so it is not in one-to-one correspondence (Zhang et al., 2024). For these methods to be compared we need to adjust the outputs of EL, so the system aggregates the recommendation entities into a single list, akin to sentence linking.

In Table 2 we present a brief summary of the comparable results discussed so far. In all these experiments, the Preferred Labels are used as retrieval options. Regarding the **Occupations**, since there exists one possible correct entity in our evaluation set, we can apply direct comparisons. Plain title similarity is the optimal strategy, where EL outperforms SL. In the case of EL and title similarity, we can observe the error propagation of the entity extraction, with a drop about 6 % accuracy. For the **Skills** and **Qualifications**, we perform the aggregation discussed. We return a list of the top-k similar entities based on the highest cosine similarity score, as with sentence similarity. In all cases, we observe EL to have a significant boost to the results.

6 Transformer Decoder Integration

In this section we explore different avenues of integrating the latest generation of LLMs for the task of linking sentences to the ESCO taxonomy.

From our experimentation, we concluded that linking job descriptions to ESCO with LLMs, like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023) directly was impossible at the time. It requires an understanding ESCO's hierarchical structures and precise concept definitions (le Vrang et al., 2014), where LLMs often produce hallucinations regarding the exact ESCO codes/ labels.

For a thorough evaluation, we opt to perform the ER task using a general-purpose decoder, the `Gemini 1.5 Pro` model and an open-source one, the `Universal-NER` (Zhou et al., 2023) model where the authors fine-tuned Llama (Touvron et al., 2023) to task-adapt it for ER and to output JSON format strings. We use the same prompt template for both models (Appendix C). We measure the performance of the models in terms of

	Method	Embedding Model	Entity Model	Accuracy@1
Occupations	Entity Linking	all-mpnet-base-v2-FT	roberta-base	0.4261
	Sentence Linking	all-mpnet-base-v2-FT	-	0.2934
	Title Linking	all-mpnet-base-v2-FT	-	**0.5387**
Skills	Entity Linking	all-mpnet-base-v2	roberta-base	**0.3969**
	Sentence Linking	all-mpnet-base-v2	-	0.2116
EQF	Entity Linking	all-mpnet-base-v2	roberta-base	**0.2881**
	Sentence Linking	all-mpnet-base-v2	-	0.1837

Table 2: **Retrieval comparison**: With **bold** we denote the best experiments for Occupation, Skill and Qualification reference sets, referred to section 2. All experiments consider only ESCO's Preferred Labels as the retrieval items.

strict F1-score, where `Gemini 1.5 Pro` achieves 0.22 with one-shot prompting and 0.25 with five-shots. `Universal-NER` reaches 0.33. Both models, severely underperform supervised methods.

Previous studies (Nguyen et al., 2024; Wang et al., 2023) have consistently shown that supervised approaches substantially outperform decoder-only models in terms of classification accuracy and consistency. These findings underscore the importance of domain-specific, fine-tuned decoders (Herandi et al., 2024) over reliance on in-context learning alone (Nguyen et al., 2024). Nonetheless, transformer-based decoders have demonstrated utility (Decorte et al., 2023; Clavié and Soulié, 2023) in re-ranking the outputs of retrieval models—an avenue not explored in the present work.

On the other hand, one of the most prominent uses of transformer decoders is their ability to create synthetic data (Clavié and Soulié, 2023). Inspired by the work of (Li et al., 2023), where they summarize the job description before performing similarity, using `Gemini 1.5 Pro`, we generate a new query from each sentence in Occupation and Skill evaluation sets. We prompt the model to produce sentences comparable to what a user with the given skill or occupation would tell the model when asked to describe their skills or occupation (Appendix C). Then, we embed such queries using `all-mpnet-base-v2` and its fine-tuned version. Detailed experiments can be found in table 10. In terms of Occupations, we observe that this method yields better results than plain sentence linking but not entity linking. For the Skills, we see a slight drop in accuracy. This indicates that synthetic query generation implements occupation matching but not skill linking.

7 Conclusion

In this study, we investigated optimal strategies for leveraging large language models (LLMs) to link job vacancy texts to the ESCO taxonomy. Emphasizing the use of open-source models, we compared two main approaches: sentence linking (SL) and entity linking (EL), with the latter incorporating an entity recognition (ER) component. Our findings indicate that EL consistently outperforms SL methods. However, we note that EL introduces greater complexity and computational overhead compared to SL. To support continued research and practical adoption, we release our entity linking system[3] and advocate for the integration of ER components into information extraction pipelines within the employment domain.

Furthermore, we introduced two novel datasets to support the evaluation of occupation and qualification extraction tasks, thereby broadening the focus beyond skill extraction for ESCO.

Given the richness of textual information in ESCO nodes, we investigated effective embedding strategies for retrieval. For Occupations, we found that combining multiple fields—preferred labels, descriptions, and concatenated alternative labels—yields the best performance. For Skills, embedding only the preferred labels proved most effective and computationally efficient.

Lastly, we achieved state-of-the-art performance on the Green Benchmark Dataset (Green et al., 2022) for entity extraction, attaining an F1 score of 54.3—surpassing the previous best of 51.2 reported by Zhang et al. (2023).

[3]To be disclosed following the review process

Limitations

Data Diversity and Language This research was done primarily on English-speaking datasets, which could limit its effectiveness in job markets with diverse linguistic profiles. Expanding handle multiple languages is recommended for future research. Additionally, the ESCO framework is designed for Europe and may not capture precisely the low- and middle-income countries' job market. Perhaps other (a few) Occupations exist in their countries that do not exist in the ESCO. In every language, and in an English setting, the specific country context has limitations, such as idioms used to refer to occupations or specially named Qualifications. There exists ongoing research regarding this topic [4].

No Joint Training The lack of a comprehensive, AIDA-style (Hoffart et al., 2011) dataset tailored for entity linking job descriptions to taxonomies like ESCO presents a significant limitation. Existing datasets fail to capture the variability and context-dependent nature of job-related terminology and they focus on different kinds of entites. This deficiency hinders the development and evaluation of robust entity linking models, particularly those designed for joint training across diverse job domains.

ESCO Node Interconnections The ESCO taxonomy includes defined links between Occupations and Skills, and within the ISCO hierarchy, interconnections also exist between various Occupations. In this work, we did not incorporate these structural relationships. However, leveraging these interconnections could potentially enhance model predictions if integrated appropriately in future implementations.

Closed-Source Models With the exception of Gemini, all models used in this study are open-source. While closed-source models have demonstrated superior performance in various scientific studies, we intentionally prioritized open-source alternatives to ensure transparency, reproducibility, and accessibility. This choice may have resulted in a trade-off in terms of maximum achievable performance.

Ethics Statement

Ethical standards were strictly adhered to throughout the research. Data collected was sourced legally and ethically from public sources, with sensitive and personally identifiable information excluded to protect privacy. The sensitive information in the original files has been redacted using Google DLP API[5]. As part of this research, we release three datasets to support transparency, reproducibility, and further investigation by the research community. These datasets are made publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The entity linking tool was designed to be fair, unbiased, and transparent. The use of large foundational models, BERT and SBERT, allows for handling various text sources. The tool's performance and results were thoroughly evaluated and documented to ensure transparency. Recognizing the tool's significant potential impact on the job market, the authors also acknowledge its limitations, such as reliance on existing data sources and potential errors or biases. Ongoing evaluation and refinement are emphasized to maintain effectiveness and fairness. Future research directions include expanding data sources, improving performance in specific segments, and integrating the tool into existing job market analysis frameworks. The authors are committed to the responsible use of the tool, ensuring its fairness, transparency, and continued improvement.

[4] https://docs.tabiya.org/overview

[5] https://cloud.google.com/sensitive-data-protection/docs/reference/rest

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

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

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

ESCO. 2024. Qualifications. https://esco.ec.europa.eu/en/classification/qualifications [Accessed at 17.05.2025].

Europass. 2024. The european qualifications framework. https://europass.europa.eu/en/europass-digital-tools/european-qualifications-framework [Accessed at 17.05.2025].

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

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

Ann-Sophie Gnehm, Eva Bühlmann, Helen Buchs, and Simon Clematide. 2022. Fine-grained extraction and classification of skill requirements in german-speaking job ads. Association for Computational Linguistics.

Thomas AF Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208. European Language Resources Association.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

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

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New Phytologist*, 11(2):37–50.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 205–215. Linköping University Electronic Press.

Hamit Kavas, Marc Serra-Vidal, and Leo Wanner. 2025. Multilingual skill extraction for job vacancy–job seeker matching in knowledge graphs. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 146–155.

Imane Khaouja, Ghita Mezzour, and Ismail Kassou. 2021. Unsupervised skill identification from job ads. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 147–151. IEEE.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann Publishers.

Martin le Vrang, Agis Papantoniou, Erika Pauwels, Pieter Fannes, Dominique Vandensteen, and Johan De Smedt. 2014. Esco: Boosting job matching in europe with semantic interoperability. *Computer*, 47(10):57–64.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Khanh Cao Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. Rethinking skill extraction in the job market domain using large language models. *arXiv preprint arXiv:2402.03832*.

Roberto Poli, Michael Healy, and Achilles Kameas. 2010. *Theory and Applications of Ontology: Computer Applications*. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf.

Rosa Stern, Benoît Sagot, and Frédéric Béchet. 2012. A joint named entity recognition and entity linking system. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 52–60. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Klim Zaporojets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, and Isabelle Augenstein. 2022. Tempel: Linking dynamically evolving and newly emerging entities. *Advances in Neural Information Processing Systems*, 35:1850–1866.

Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022a. Skillspan: Hard and soft skill extraction from english job postings.

Mike Zhang, Kristian Nørgaard Jensen, Rob van der Goot, and Barbara Plank. 2022b. Skill extraction from job postings using weak supervision.

Mike Zhang, Rob Van Der Goot, and Barbara Plank. 2023. Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Entity linking in the job market domain. *arXiv preprint arXiv:2401.17979*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

Appendix

A Training Hyperparameters

In Table 3, we present the model parameters that were used during this research.

Model Sizes	Parameters
roberta-base	124M
roberta-large	354M
bert-base-cased	107M
bert-large-cased	332M
deberta	138M
all-mpnet-base-v2	104M
all-MiniLM-L6-v2	22M

Table 3: Model Sizes and Parameters

A.1 Entity Recognition

For the ER training we did a hyperparameter search regarding the parameters batch size, epochs, and learning rate. We set the max length of the sentences to 128 tokens weight decay to 0.01, while we search from possible options batch size:16, 32, 64, epochs:5, 10, learning rate: 0.0001, 0.00005, 0.00001. In Table 4, we present the best hyperparameters with respect to the results in Table 2. ER evaluation was performed with three random seed initialization 3, 37 and 42 on the HuggingFace token classification script.

model	batch	lr	epoch
bert-base	16	5e-5	10
bert-base+CRF	32	1e-4	5
bert-large	64	1e-4	10
bert-large+CRF	16	5e-5	10
deberta-base	32	5e-5	5
deberta-base+CRF	32	5e-5	5
jobbert	32	5e-5	10
jobbert+CRF	32	1e-4	5
roberta	32	1e-4	5
roberta+CRF	16	1e-4	5
roberta-large	32	5e-5	5
roberta-large+CRF	32	5e-5	5
ESCOXLM-R	32	5e-5	5
ESCOXLM-R+CRF	32	5e-5	5

Table 4: Entity Recognition best training hyperparameters

Parameter	Value
Epochs	2
Evaluation Steps	0
Evaluator	NoneType
Max Gradient Norm	1
Optimizer Class	AdamW
Learning Rate (lr)	2e-05
Scheduler	WarmupLinear
Warmup Steps	10000
Weight Decay	0.01

Table 5: Summary of the training configuration parameters for Sentence Transformers

A.2 Entity Similarity

For the entity similarity training, we used the sbert official website to train our models. We resulted on training with the default parameters without hyperparmeter search. In Table 5, we present the hyperparameters.

B Dataset tables

All analysis in this section was done with the NLTK [6] package.

		Statistics
TRAIN	Sentences	9,634
	Tokens	233,628
	Entity Spans	18,098
TEST	Sentences	336
	Tokens	8,050
	Entity Spans	904
	Average Entity Length	3.67

Table 6: **Green Benchmark** Data Analysis

	Statistics
# of pairs	210,175
# ESCO occupations	1,156

Table 7: **Ethiopian Jobs Training Set** Data Analysis

C Prompts used in this study

The following prompts have been used in section 2 to judge between the qualification annotations.

[6] https://www.nltk.org/

EQF Level	Statistics
1	40
2	88
3	89
4	166
5	115
6	128
7	117
8	74
Total	814
Average Word Length	7.24
Total Countries	30
Entries per country	27.13

Table 8: **EQF reference database** Data Analysis

Statistics	Occupations	Skills	EQF
Data points	542	920	448
Avg entities	1	2.7	1.3
Avg words	418.2	16.5	29
Entities	542	2406	595
Words per entity	3.4	3.1	3.4
Max entities	1	31	7
Number of UNK	0	981	361

Table 9: **Evaluation Sets** Data Analysis with NLTK

Prompt: "In the context of the following sentence choose the appropriate EQF level that suits the qualification. If you cannot determine the EQF level answer UNK.
Example:
Sentence: Qualifications and experiences : BSc , MSc or PhD or equivalent in Statistics , Computer Science , Mathematics or other analytical field .
Qualification: BSc , MSc or PhD or equivalent in Statistics , Computer Science , Mathematics or other analytical field .
EQF level: EQF8
Sentence: sentence
Qualification: qualification
EQF level: "

The following prompts have been used in section 6 to generate synthetic queries from Occupations and Skills datasets for evaluation. For each datapoint, the prompt is adapted depending on the original job title, job description or skill description.

Occupation prompt: "Given the following description of the user's past job, return the answer of the user to the following question.

Description: <*title*> <*description*>

Question: Describe your last job. Answer in one sentence. Don't be too formal.

Answer:"

Skill prompt: "Given the following description of the user's skill, return the answer of the user to the following question. Description: <*description*> Question: What are your skills and expertise? Answer in one sentence. Don't be too formal. Answer:"

Lastly we present the prompt template used in section 6 to perform entity extraction with transformer decoders.

Prompt: "A virtual assistant answers questions from a user based on the provided text.
$few shots$
USER: Text: $text$
ASSISTANT: I've read this text.
USER: What describes $entity$ in the text?
ASSISTANT: "
where we replace the few shot, text and entity placeholders with data points from the Green Benchmark.

D Sentence Linking results

In this section, we present the analytical results of the experiments mentioned in section 3 and section 6. Table 10 denotes the initial vector search where we embed full sentences and table 11 the title similarity experiments.

		Fulltext		Synthetic query	
	Embeddings selection	**mpnet**	**mpnet-ft**	**mpnet**	**mpnet-ft**
Occupations	Single: Preferred Label	0.1974	0.2934	0.2177	0.3506
	Single: Description	0.2657	0.3745	0.2675	0.4022
	Single: Preferred Label and Description	0.2915	0.3635	0.2878	0.4133
	Single: All fields	0.2454	0.3450	0.2749	0.3616
	Multiple: All fields	0.2874	**0.3945**	0.3415	**0.4256**
	Multiple: All fields separated	0.2612	0.3542	0.2884	0.3965
Skills	Single: Preferred Label	0.2116	0.1678	0.1783	0.1573
	Single: Description	0.1211	0.081	0.0858	0.0705
	Single: Preferred Label and Description	0.2059	0.1554	0.1401	0.1411
	Single: All fields	**0.2212**	0.1697	**0.2050**	0.1582
	Multiple: All fields	0.1554	0.1487	0.1386	0.1311
	Multiple: All fields separated	0.1516	0.1430	0.1335	0.1286
EQF	Single	0.1837	0.1880	-	-

Table 10: **RAG-related Vector Search: sentence linking** The best results on each experiment are denoted in bold. Single indicates that only one embedding was generated for each target ESCO node, while multiple indicates than more than one embedding was generated. It is important to note that this evaluation is done on the sentence level. The fulltext column refers to experiments done in section 3 and the synthetic query to the experiments of section 6.

Occupations	Single: Preferred Label	0.3764	0.5387
	Single: Description	0.3339	0.4686
	Single: Preferred Label and Description (concatenated)	0.3339	0.4502
	Single: All fields (concatenated)	0.3321	0.4446
	Multiple: All fields (separated)	0.3782	**0.5406**
	Multiple: All fields (separated secondary labels)	0.3321	0.5018

Table 11: **RAG-related Vector Search: title linking** Experiment for occupation linking using job titles as queries