Enhancing Job Posting Classification with Multilingual Embeddings and Large Language Models

Hamit Kavas[1,2,*], Marc Serra-Vidal[2] and Leo Wanner[1,3]

[1]NLP Group, Pompeu Fabra University, C/ Roc Boronat, 138, 08018, Spain
[2]Adevinta Spain, C/ de la Ciutat de Granada, 150, Barcelona, 08018, Spain
[3]Catalan Institute for Research and Advanced Studies (ICREA), Passeig Lluís Companys, 23, Barcelona, 08010, Spain

Abstract

In the modern labour market, taxonomies such the *European Skills, Competences, Qualifications and Occupations* (ESCO) classification are used as an *interlingua* to match job postings with job seeker profiles. Both are classified with respect to ESCO *occupations*, and match if they align with the same occupation and the same skills assigned to the occupation. However, matching models usually struggle with the classification because of overlapping skills and similar definitions of occupations defined in the ESCO taxonomy. This often leads to imprecise classification outcomes. In this paper, we focus on the challenge of the classification of job postings written in Italian or Spanish against ESCO occupations written in English. We experiment with multilingual embeddings, zero-shot classification, and use of a large language model (LLM) and show that the use of an LLM leads to best results. Furthermore, we also explore an alternative automatic labeling method by prompting three top-performing LLMs to annotate the test dataset. This approach serves both as an experiment on the usability of automatic labeling and as an evaluation of the reliability of the automatically assigned labels, involving human annotators.

Keywords

ESCO labour market taxonomy, job posting classification, class embeddings, text embeddings, LLM

1. Introduction

The modern labour market becomes more and more diverse. High-tech jobs demand novel skills and competences, which in their turn keep undergoing adaptations and modifications. Under these circumstances, accurately classifying job postings and CVs of job seekers (henceforth *candidate experiences*) that contain detailed technological specifications with remarkably similar yet distinct skills and experiences has evolved into a complex challenge.

The overwhelming majority of job portals and employment agencies use either the *European Skills, Competences, Qualifications and Occupations* (ESCO) taxonomy[1] or its US equivalent O*Net taxonomy[2] to classify job postings and candidate experiences in terms of job title labeled ESCO/O*Net occupations. Most of the proposals to automatic alignment of job postings with candidate experiences (or vice versa) also use ESCO or O*Net [1, 2, 3]. However, despite their wide use, both ESCO and O*Net taxonomies exhibit principle limitations for the task of automatic classification of job postings and candidate experiences because due to their tree structure they often fail to adequately distinguish between occupations that exhibit substantial skill overlaps. For instance, two job postings labeled as 'data analyst' may appear similar but require different skills if one focuses on market research while the other concentrates on healthcare trends analysis. This issue is particularly pronounced when classification relies on a single label, such as the job title of an ESCO occupation, where skill overlaps undermine precise classification. Hence, employing multiple job titles and framing the problem as a multi-label classification task is imperative.

This paper addresses the challenge of multilingual multi-label classification using Large Language Models (LLMs) for the alignment of Italian and Spanish job postings with English job titles encountered in the ESCO taxonomy. Multilingual class embeddings are explored to improve classification accuracy, aiming to provide the necessary contextual awareness and addressing the core limitations of taxonomies such as ESCO.

Furthermore, we explore an alternative automatic labeling method by prompting three top-performing LLMs to annotate the test dataset. This approach serves both as an experiment on the usability of automatic labeling and as an evaluation of the reliability of the automatically assigned labels, involving human annotators.

To provide LLMs with domain-specific information and to mitigate hallucinations in the course of the classification of the job postings, we employ Retrieval Augmented Generation (RAG) [4], which combines information retrieval with a generative model. RAG serves

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

hamit.kavas@upf.edu (H. Kavas); marc.serrav@adevinta.com (M. Serra-Vidal); leo.wanner@upf.edu (L. Wanner)
0009-0009-7027-7367 (H. Kavas); 0009-0000-0120-381X (M. Serra-Vidal); 0000-0002-9446-3748 (L. Wanner)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1] https://esco.ec.europa.eu/en/classification

[2] https://www.onetonline.org/

CEUR-WS.org/Vol-3878/52_main_long.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

two critical functions in our methodology. Firstly, it provides detailed definitions, including essential skills and synonyms for each ESCO occupation, selected through vector similarity as outlined in [5]. Secondly, it ensures that the assigned job titles are restricted to titles within our predefined label space, i.e., standardized job titles defined in the ESCO taxonomy.

The contributions of our work are:

• We explore the impact of using multilingual class embeddings derived from the ESCO taxonomy for the task of job posting classification.
• We integrate RAG to provide LLMs with domain-specific information and eliminate the dependency on fine-tuning;
• We show how the LLM response can be restricted to standardized job titles and thus how LLMs can be used for high quality job title classification that outperforms state-of-the-art proposals for this task.

The remainder of the paper is structured as follows. In Section 2, we present a concise overview of the related work. In Section 3, the model on which our work is based is outlined. Section 4 describes the experiments we carried out, the results we obtained in these experiments, and their discussion. In Section 5, finally, draws some conclusions from the presented work and outlines some directions for future research. In Appendix A, we present an ablation study in which we assess the comprehension of English ESCO job titles and its Spanish equivalents by our model. Appendix B provides, for illustration, examples of Italian job postings and predicted ESCO job titles. In Appendix E, we present the signature used to prompt Large Language Models for pre-processing.

2. Related Work

A number of works have been carried out in the domain of job title classification, focusing on various facets of the problem. Shi et al. [6] introduce Job2Skills, a model developed for LinkedIn. The model significantly improves job recommendation performance metrics, however, raises questions about its effectiveness beyond LinkedIn. Li et al. [7] proposes a two-step job title normalization, also in LinkedIn, which is based on tokenization and matching of the original job title provided by the user with a lookup table. The use of a lookup table instead of a standard occupation taxonomy such as ESCO or O*Net significantly limits the generalization potential of this strategy. Zhang et al. [8] extract soft and hard skills from job posting descriptions, showing that domain-specific pre-training significantly enhances performance in skills and knowledge extraction. Javed et al. [3] introduce a semi-supervised machine learning approach that utilizes hierarchical classifiers and the O*NET *Standard Occupational Classification* (SOC) taxonomy for the classification of online recruitment data. Similarly, Wang et al. [9] propose a model based on multi-stream convolutional neural networks, aiming to classify noisy user-generated job titles by considering different elements such as characters and words within job titles. Yamashita et al. [10] and Zbib et al. [1] conduct studies on the classification of job titles, focusing on job title alignment and job similarity training, respectively. JobBERT Decorte et al. [2] classifies job titles against the ESCO taxonomy, treating the task as a semantic text similarity (STS) exercise. In particular, JobBERT emphasizes the understanding of the semantics of job titles through the skills inferred from the associated vacancies and descriptions, thus alleviating the need for an extensive labeled dataset or a continuously updated list of standardized titles. Before the recent proposals [11] and [12], JobBERT used to be referenced as the state-of-the-art baseline. In general, all of these works draw upon some of the information encoded in the ESCO taxonomy. However, none of them uses detailed descriptions of ESCO occupations, as we propose.

3. The Model

3.1. The Basics

The proposed model is based on the notion of *distinctiveness*, which specifies the difference between the prompt concept θ^* and other concepts within the conceptual space Θ [13]. The notion is crucial for distinguishing in-context learning concepts that are aimed to be learned by analogy. θ^* acts as a latent parameter in a Hidden Markov Model that defines a distribution over observed tokens, represented by selected ESCO job titles as labels. As proposed by Xie et al. [13], the error of the in-context predictor approaches optimality under the condition that θ^* is distinguishable from other concepts in $\Theta \setminus \{\theta^*\}$. When RAG is adapted as a few-shot reasoning (or in-context learning) framework for job posting classification [14] , θ^* is represented by the top-selected ESCO labels and ensures that the LLM can effectively differentiate between closely related job categories.

The explanation enriched prompts enhance the LLM's ability to learn more from each example. According to Xie et al. [13], the expected error decreases as the length and informational content of each example increase, contributing to the richness of the input–output mapping for a more robust in-context learning environment. This assumption is proven to be true under the condition of distinguishability of in-context examples and can be mathematically expressed as a reduction of the expected error $E[\epsilon]$, correlated with an increase in the information content I of the examples:

$$E[\epsilon] \propto \frac{1}{I(S_n, x_{\text{test}})} \tag{1}$$

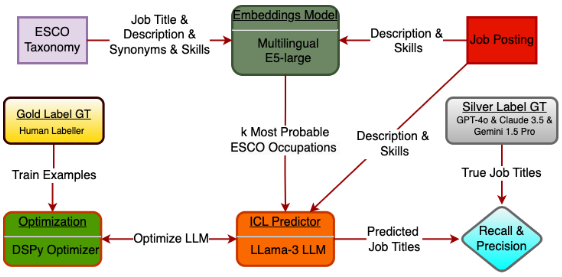
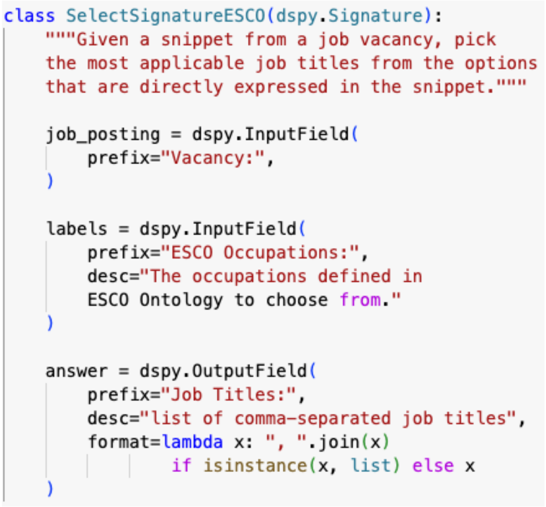


Figure 1: Model Architecture



```
class SelectSignatureESCO(dspy.Signature):
    """Given a snippet from a job vacancy, pick
    the most applicable job titles from the options
    that are directly expressed in the snippet."""

    job_posting = dspy.InputField(
        prefix="Vacancy:",
    )

    labels = dspy.InputField(
        prefix="ESCO Occupations:",
        desc="The occupations defined in
        ESCO Ontology to choose from."
    )

    answer = dspy.OutputField(
        prefix="Job Titles:",
        desc="list of comma-separated job titles",
        format=lambda x: ", ".join(x)
                if isinstance(x, list) else x
    )
```

Figure 2: Prompt Template

where S_n represents the sequence of training examples in the prompt and x_{test} is the test input.

The use of RAG helps avoid hallucination since when directly prompted with job postings, LLMs have been observed to sometimes produce non-existent labels [5].

3.2. Design of the Model

The proposed model (see Figure 1) uses multilingual class embeddings of the E5-large model[15] to retrieve pertinent ESCO occupation definitions in English. The definitions serve as contextual information to prompt language models for selection of the most suitable job titles. To this end, we incorporate the DSPy library's Chain-of-Thought mechanism,[3] augmented by a hint to restrict the model output to a specified list of job titles. The signature used in this methodology (cf. Figure 2) is inspired by [16].

To implement the RAG model, we initially established a vector database,[4] in which English ESCO occupation definitions were inserted as multilingual embedding vectors. Acknowledging the reported significance of chunking in many NLP applications, we conducted a series of ablation studies to determine the optimal chunk size. These studies revealed that subdividing the ESCO occupation definitions into smaller segments adversely affects the performance of vector-based similarity matching. Therefore, we opted for storing each of the 3,015 occupations represented in the ESCO taxonomy in its entirety.

Table 1
Recall values for classification with E5-large Text Embeddings vector similarity

Precision @ K	@5	@10	@30	@40
Value	0.4238	0.9004	0.9627	0.9817

To accurately classify a given job posting with respect to the ESCO taxonomy, we include 30 ESCO occupation documents (i.e., 30 nodes of the taxonomy) into the LLM's context as potential job titles. The rationale for choosing 30 documents is that we aim to strike a balance between computational efficiency and the accuracy of the retrieved documents. The precision of the LLM would naturally decrease when it is presented with inaccurate labels. Although, as shown in Table 1, the precision of the model slightly increases with 40 documents in the context, we accepted this trade-off in favor of a lower VRAM requirement.

Upon the retrieval of the 30 ESCO occupations that are most closely aligned with a given job posting description, a composite prompt (see Figure 2) is constructed as input to the LLM. The prompt integrates the actual text data encompassing job titles, descriptions, and skills pertinent to the selected occupations. The design of the simplified composite prompt aims to minimize the bias by focusing only on the core elements. The prompt is then processed by using a locally stored Llama-3 LLM[5] in an isolated environment[6].

As a few-show predictor, the LLM evaluates the composite prompt to accurately classify job postings by examining the semantic nuances of the selected ESCO occupations, aligning them with the actual job titles within the offers. To quantitatively assess the alignment between a job posting vector J and each occupation embedding E_{ESCO} derived from the ESCO taxonomy, cosine similarity $a(J, E_{\text{ESCO}})$ is used:

$$a(J, E_{\text{ESCO}}) = \frac{J \cdot E_{\text{ESCO}}}{\|J\| \|E_{\text{ESCO}}\|} \tag{2}$$

The similarity scores yielded through $a(J, E_{\text{ESCO}})$ for each E_{ESCO} facilitate the identification and selection of

[3] https://github.com/stanfordnlp/dspy
[4] https://www.trychroma.com/
[5] https://llama.meta.com/llama3/
[6] We use dockerized models from the open-source Ollama library https://ollama.com/ for all experiments

the ESCO occupation embeddings that are most pertinent to the job posting in question. Armed with this information, the LLM proceeds to classify the job posting by selecting the ESCO occupation that exhibits the highest degree of semantic and contextual relevance.

For a specific job posting J, an embedding function E is employed, such that $E(J)$ produces the corresponding embedding for J. The degree of similarity between the job posting's embedding $E(J)$ and any ESCO occupation embedding e_i from E_{ESCO} (where E_{ESCO} stands for the ensemble of occupation embeddings derived from the ESCO taxonomy) is determined through the similarity function $S(E(J), e_i)$ (in our case cosine).

The similarity scores for each occupation embedding e_i within E_{ESCO} relative to $E(J)$ are computed. The ten class embeddings that exhibit the highest similarity to $E(J)$, denoted as E_{top}, are selected. Formally, E_{top} is defined as the subset $\{e_1, e_2, \ldots, e_{10}\}$ from E_{ESCO}, where each e_i is selected based on the top 10 similarity scores $S(E(J), e_i)$.

The last stage entails a decision-making process enacted by the Llama-3 LLM, represented by the function D. This function accepts the composite prompt including candidates $\{e_1, e_2, \ldots, e_{10}\}$ accumulated to E_{top} and the job posting J, to render the final selected occupation embedding. The chosen occupation embedding e^* is determined by $e^* = D(E_{\text{top}}, J)$, representing the ESCO occupation best matched by the model.

The entire algorithm can be presented by the following equation, which encapsulates the embedding generation, similarity assessment, and decision-making process by the LLM, culminating in the selection of the most suitable ESCO occupation embedding e^* for the given job posting description.

$$e^* = D(\{e_1, e_2, \ldots, e_k \mid e_i \in E_{\text{ESCO}}; \\ \text{top k by } S(E(J), e_i)\}, J) \quad (3)$$

4. Experiments

To evaluate the effectiveness of the proposed model in handling multilingual job postings, experiments were conducted separately on Italian and Spanish datasets.

4.1. Test dataset

To have a reliable test dataset, we use three high performing LLMs as initial annotators of real-world 100 Italian and 100 Spanish job postings with the most extensive descriptions from the InfoJobs [7] database. Non-informative elements such as company descriptions and promotional content where removed using a DSPy module (cf. Appendix E for prompt), which employs zero-shot LLama-3 LM inference to anonymize sensitive information in job postings and candidate experiences. The preprocessed postings were annotated by the top three performing LLMs: GPT-4o[8], Gemini 1.5 Pro[9], and Claude 3.5 Sonnet[10], according to LmSys Arena[17]. In this context, the ESCO job titles are presented to each model separately, requesting them to select the appropriate job titles, and then measure their level of agreement on these labels. The agreement between LLM models was assessed using Cohen's kappa coefficient[18]. The average kappa score between Gemini and GPT-4o was found to be 0.6386, indicating a substantial level of agreement. The agreement between Gemini and Claude was lower, with an average kappa of 0.5798, suggesting a moderate level of agreement. Similarly, the kappa score between GPT-4o and Claude was 0.6497, also indicative of substantial agreement. Overall, the average kappa score across all "annotators" was 0.6227, reflecting a general trend towards substantial inter-annotator agreement among the models.

To establish ground truth labels, we incorporated a dual-layer labelling process. Although the test set consists of only 200 items, labeling them from scratch would be time-consuming due to the complexity of the ESCO taxonomy, which includes 3,015 distinct occupations. Human annotators would require extensive training to accurately navigate this taxonomy. Therefore, we first annotate the occupations automatically using LLMs and then let the initial annotations cross-examine by human expert annotator. Since each data point was reviewed by one annotator only, inter-annotator agreement among human annotators was not quantified. Instead, we conducted an analysis to identify job titles that consistently showed agreement or disagreement across the three LLMs, where domain-specific professionals from InfoJobs reviewed label discrepancies. This analysis, detailed in Appendix C, suggests that certain occupations are inherently more challenging to classify, possibly due to overlapping skills or ambiguous descriptions.

Furthermore, we repeated experiments using ground truth labels where any two of the three automatic LLMs agreed on the label. The results showed alignment between the models' predictions and the automatic labeling process, indicating consistency with the patterns recognized by the automatic methods when there is partial agreement. A detailed analysis of this alignment can be found in Appendix D.

[7] https://www.infojobs.net/

[8] https://openai.com/index/hello-gpt-4o/

[9] https://deepmind.google/technologies/gemini/pro/

[10] https://www.anthropic.com/news/claude-3-5-sonnet

4.2. Baselines

4.2.1. SkillGPT

SkillGPT [5] has been introduced as a tool for skill extraction and classification, with vector similarity search against LLM-precomputed ESCO embeddings. The authors employ embeddings generated by an LLM, although they do not directly use LLM to select among candidate embeddings. Instead, they rely on embedding similarity to assign the most closely related ESCO class to job descriptions under consideration.

4.2.2. Zero-Shot Classification

By transforming the classification task into a Natural Language Inference (NLI) problem, any model pretrained on NLI tasks can be utilized as a text classifier without the need for fine-tuning, effectively achieving zero-shot text classification. This is particularly beneficial when we deal with classes unseen during training, making it a robust solution for a variety of text classification scenarios [19].

In our implementation that we use as baseline, we utilize the BART-MNLI model [20] that showed high performance in summarization tasks when pretrained for various NLI tasks on an MNLI dataset [21] that is leveraged for its capability to understand entailment relations for classification of the given sequence into one of the specified categories. We also apply the same methodology with the Llama-3 model.

4.3. Model Optimization

To optimize LLMs with a minimal set of manually crafted examples, we use the DSPy library [22]. We initialize the classifier module with a Llama-3 model and use a GPT-4o model as the teacher. Our optimization of the classification is aimed at achieving high F1 scores for each dataset individually. In each run, we use 10 labeled training examples and 30 labeled validation examples. We employ DSPy's *BootstrapFewShot*, configuring it to perform a maximum of 2 rounds with up to 8 bootstrapped demonstrations. We define a custom metric—the F1 score—to guide the bootstrapping process. For the optimization of the LLMs, we use data points that had high inter-agreement among the automatic methods and were reviewed by human annotators. We perform a validation/test split to ensure that the optimization did not bias the evaluation results.

4.4. Outcome of the experiments

For the evaluation of the results of the experiments, we used the *micro recall* and *micro precision* metrics, which are suitable for our multi-class classification task. We report evaluation scores seperately on Spanish and Italian test sets.

Table 2
Italian Performance Metrics for Top 5 and Top 10 Predictions

Model	Precision		Recall	
	@5	@10	@5	@10
llama-3-8b (CoT opt.)	**0.32**	0.13	**0.76**	0.80
llama-3-8b (CoT)	0.26	0.12	0.62	0.64
llama-3-8b (SkillGPT)	0.19	0.19	0.36	0.82
mBart-large-mnli (0-shot)	0.13	0.12	0.29	0.58
multilingual-e5-large	0.16	0.19	0.36	0.88

Table 3
Spanish Performance Metrics for Top 5 and Top 10 Predictions

Model	Precision		Recall	
	@5	@10	@5	@10
llama-3-8b (CoT opt.)	0.28	**0.20**	0.72	**0.90**
llama-3-8b (CoT)	0.26	0.16	0.64	0.68
llama-3-8b (SkillGPT)	0.09	0.12	0.36	0.62
mBart-large-mnli (0-shot)	0.15	0.14	0.39	0.70
multilingual-e5-large	0.20	0.19	0.48	**0.92**

Tables 2 and 3 display the results on the Italian and Spanish datasets, respectively. The results indicate that prompting techniques outperform SkillGPT in both languages. Specifically, the optimized Llama-3-8b model with chain-of-thought (CoT) achieves the highest precision and recall at @5 for Italian, with values of 0.32 and 0.76, respectively, and for Spanish, with values of 0.28 and 0.72. This supports our assumption that optimization enhances performance. The multilingual E5-large model achieves the highest precision at @10 for Italian (0.19) and the highest recall at @10 for Spanish (0.92), underscoring the efficacy of embeddings in classification. This implies that semantically less similar labels can confuse models, whereas embeddings ensure higher recall accuracy, particularly in wider retrieval scenarios. Although both models exhibit similar precision, indicating comparable accuracy in their predictions, the optimized model's capacity to capture a broader range of relevant job titles ensures greater alignment with expert human preferences. This enhances the model's ability to make relevant job title suggestions, thereby improving the overall matching process.

4.5. Discussion

In Tables 2 and3 we observe that the combined use of general text embeddings and language models significantly outperforms current classification techniques, which rely on language models specifically tailored to the field of the labour market, such as [12]. We see that using vector similarity with the text embeddings created by the E5-large text embedings model alone does not surpass the baseline. However, it is worth noting that the results are quite close, despite the fact that this model was not specifically fine-tuned on labour market data or adapted to the ESCO taxonomy, as is the case of [12]. Furthermore, we can observe how text embeddings indeed provide a significant value for filtering n occupations closest to a job posting within the taxonomy. Using these k professions as input to various language models for few-shot classification significantly improves over the baselines. Table 6 in the Appendix illustrates the decisions of the LLMs in the case of four sample job postings.

We also evaluated the effectiveness of a large language model for classification of job titles based on provided descriptions, as shown in Table 4 even when the correct titles were not explicitly listed among the initial ESCO job titles. The model's ability to select accurate titles reflects its functionality in processing and understanding the contextual and semantic aspects of the job descriptions. For instance, when presented with a job description focused on the management of comprehensive water and wastewater services, the model correctly identified "Operations Manager" as the correct title. This identification was made despite the presence of several closely related but distinct labels (such as, "Water treatment plant manager") within the pool of ESCO job titles. This indicates that the model's decisions are more influenced by a comprehensive understanding of the job responsibilities and sectors than by the mere presence of keywords or phrases in the ESCO job titles.

The model's capacity to differentiate between job titles with more specific definitions enhances its comprehension of job postings and assigned labels, thereby improving the precision of suggesting relevant skills. Upon integration into an operational job platform, this model will better understand the requirements of job postings and accurately assign job titles that align with the specific needs of companies. Similarly, in the context of parsing of job candidate experiences, keywords tend to appear more frequently in semantically related ESCO definitions, enabling parsers to incorporate these keywords to enhance parsing performance.

Overall, we can thus state that the integration of class embeddings generated using the multilingual E5-large model, with subsequent application of few-shot classification techniques through LLMs, significantly improves the accuracy of job title classification, clearly surpassing those of the baselines.

4.6. Computational Cost of Compared Methods

In addition to evaluating performance metrics, we analyzed the computational cost and environmental impact of each method. The *Llama-3-8b* model, with 8 billion parameters, requires significant resources for inference, necessitating a GPU with at least 16 GB of VRAM (e.g., NVIDIA RTX 3090). Its average inference time per job posting is approximately 1.5 seconds, and its high energy consumption leads to increased CO_2 emissions, making large-scale deployment less environmentally sustainable without optimizations.

In contrast, the *mBART-large-mnli* model has about 610 million parameters and operates on GPUs with 8 GB of VRAM, offering faster inference times under 0.5 seconds per job posting. The embeddings-based method using the *multilingual E5-large* model, with 330 million parameters, allows for precomputed embeddings and efficient CPU-based vector similarity searches, reducing inference time to less than 0.2 seconds per job posting. These smaller models consume less energy, providing more resource-efficient and eco-friendly alternatives suitable for production environments where computational cost and environmental impact are critical considerations.

5. Conclusions and future work

In this paper, we argued that the use of multilingual embeddings in combination with LLMs significantly enhances our ability to distinguish between very similar (or even identical) job titles that suggest different skills and competencies. Our experiments have shown that this is indeed the case, demonstrating that the combination of multilingual text embeddings similarity with the Llama-3 markedly exceeds the performance of other leading approaches in the field.

In the future, we plan to apply the same approach to the analysis and classification of job candidate experiences. Once it is ensured that both job postings and candidate experiences can accurately be modeled using the embedded representation of the ESCO taxonomy, we plan to set the stage for a more direct and efficient alignment process between job postings and experiences of job seekers.

Another interesting direction for future research is to analyze the lexical overlap between English domain-specific terms that appear in Italian and Spanish job postings and the English occupation descriptions in the ESCO taxonomy. Such an analysis would reveal whether job types with higher lexical overlap affect model accuracy, providing deeper insights into the multilingual nature of the task.

References

[1] R. Zbib, L. L. Alvarez, F. Retyk, R. Poves, J. Aizpuru, H. Fabregat, V. Šimkus, E. G. Casademont, Learning job titles similarity from noisy skill labels, ArXiv abs/2207.00494 (2022). URL: https://api.semanticscholar.org/CorpusID:250243975.

[2] J.-J. Decorte, J. V. Hautte, T. Demeester, C. Develder, Jobbert: Understanding job titles through skills, ArXiv abs/2109.09605 (2021). URL: https://api.semanticscholar.org/CorpusID:237572142.

[3] F. Javed, M. McNair, F. Jacob, M. Zhao, Towards a job title classification system, 2016. arXiv:1606.00917.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[5] N. Li, B. Kang, T. D. Bie, Skillgpt: a restful api service for skill extraction and standardization using a large language model, 2023. arXiv:2304.11060.

[6] B. Shi, J. Yang, F. Guo, Q. He, Salience and market-aware skill extraction for job targeting, 2020. arXiv:2005.13094.

[7] S. Li, B. Shi, J. Yang, J. Yan, S. Wang, F. Chen, Q. He, Deep job understanding at linkedin, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2020. URL: http://dx.doi.org/10.1145/3397271.3401403. doi:10.1145/3397271.3401403.

[8] M. Zhang, K. N. Jensen, S. D. Sonniks, B. Plank, Skillspan: Hard and soft skill extraction from english job postings, ArXiv abs/2204.12811 (2022). URL: https://api.semanticscholar.org/CorpusID:248405777.

[9] J. Wang, K. Abdelfatah, M. Korayem, J. Balaji, Deepcarotene -job title classification with multi-stream convolutional neural network, 2019, pp. 1953–1961. doi:10.1109/BigData47090.2019.9005673.

[10] M. Yamashita, J. T. Shen, H. Ekhtiari, T. Tran, D. Lee, James: Job title mapping with multi-aspect embeddings and reasoning, 2022. arXiv:2202.10739.

[11] M. Zhang, R. van der Goot, B. Plank, Escoxlmr: Multilingual taxonomy-driven pre-training for the job market domain, in: Annual Meeting of the Association for Computational Linguistics, 2023. URL: https://api.semanticscholar.org/CorpusID:258832782.

[12] H. Kavas, M. Serra-vidal, L. Wanner, Job offer and applicant cv classification using rich information from a labour market taxonomy, SSRN Electronic Journal (2023). doi:10.2139/ssrn.4519766.

[13] S. M. Xie, A. Raghunathan, P. Liang, T. Ma, An explanation of in-context learning as implicit bayesian inference, 2022. arXiv:2111.02080.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997.

[15] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards General Text Embeddings with Multi-stage Contrastive Learning, arXiv e-prints (2023) arXiv:2308.03281. doi:10.48550/arXiv.2308.03281. arXiv:2308.03281.

[16] K. D'Oosterlinck, O. Khattab, F. Remy, T. Demeester, C. Develder, C. Potts, In-context learning for extreme multi-label classification, ArXiv abs/2401.12178 (2024). URL: https://api.semanticscholar.org/CorpusID:267068618.

[17] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, I. Stoica, Chatbot arena: An open platform for evaluating llms by human preference, ArXiv abs/2403.04132 (2024). URL: https://api.semanticscholar.org/CorpusID:268264163.

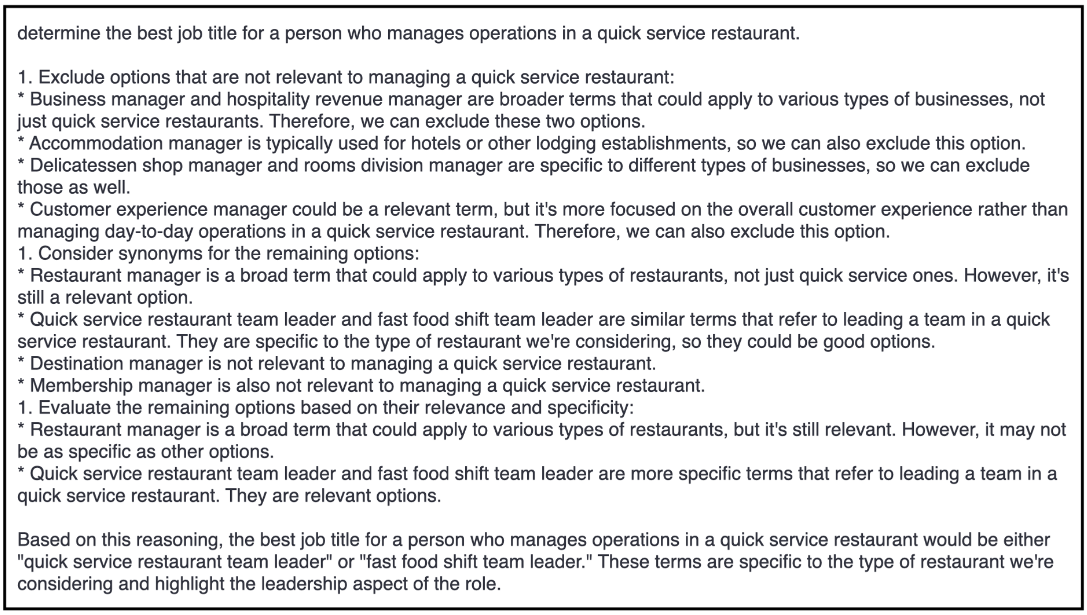
[18] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37 – 46. URL: https://api.semanticscholar.org/CorpusID:15926286.

[19] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning based text classification: A comprehensive review, CoRR abs/2004.03705 (2020). URL: https://arxiv.org/abs/2004.03705. arXiv:2004.03705.

[20] L. Shu, J. Chen, B. Liu, H. Xu, Zero-shot aspect-based sentiment analysis, ArXiv abs/2202.01924 (2022).

[21] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101.

[22] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, Dspy: Compiling declarative language model calls into self-improving pipelines, ArXiv abs/2310.03714 (2023). URL: https://api.semanticscholar.org/CorpusID:263671701.



determine the best job title for a person who manages operations in a quick service restaurant.

1. Exclude options that are not relevant to managing a quick service restaurant:
* Business manager and hospitality revenue manager are broader terms that could apply to various types of businesses, not just quick service restaurants. Therefore, we can exclude these two options.
* Accommodation manager is typically used for hotels or other lodging establishments, so we can also exclude this option.
* Delicatessen shop manager and rooms division manager are specific to different types of businesses, so we can exclude those as well.
* Customer experience manager could be a relevant term, but it's more focused on the overall customer experience rather than managing day-to-day operations in a quick service restaurant. Therefore, we can also exclude this option.
1. Consider synonyms for the remaining options:
* Restaurant manager is a broad term that could apply to various types of restaurants, not just quick service ones. However, it's still a relevant option.
* Quick service restaurant team leader and fast food shift team leader are similar terms that refer to leading a team in a quick service restaurant. They are specific to the type of restaurant we're considering, so they could be good options.
* Destination manager is not relevant to managing a quick service restaurant.
* Membership manager is also not relevant to managing a quick service restaurant.
1. Evaluate the remaining options based on their relevance and specificity:
* Restaurant manager is a broad term that could apply to various types of restaurants, but it's still relevant. However, it may not be as specific as other options.
* Quick service restaurant team leader and fast food shift team leader are more specific terms that refer to leading a team in a quick service restaurant. They are relevant options.

Based on this reasoning, the best job title for a person who manages operations in a quick service restaurant would be either "quick service restaurant team leader" or "fast food shift team leader." These terms are specific to the type of restaurant we're considering and highlight the leadership aspect of the role.

Figure 3: LLM's Rationale

A. Ablation Study

In our ablation study, we pursued two primary objectives. Firstly, to evaluate the model's comprehension of ESCO job titles and its decision-making process. To achieve this, we prompted the model to articulate its underlying rationale. Secondly, so far we reported the performance of our model when Italian and Spanish data were matched against English job titles and occupations in the ESCO taxonomy. Here we wanted to explore whether its comprehension was extendable to data in different languages. We selected Spanish for this purpose and discovered that the model's understanding was consistent, irrespective of the language; see Table 4.

As illustrated in Figure 3, the LLM showcases a comprehensive understanding of the task at hand, effectively narrowing down potential ESCO job titles to identify the most suitable label. Additionally, the LLM is observed to generate a novel job title, referred to as "fast food shift team leader". This can be attributed to the absence of contstraints imposed on the LLM regarding structured output for classification, thereby granting it to autonomy to propose the most fitting job title. The analysis initially excludes broader or less related job titles such as "bussiness manager", "hospitality revenue manager", and "accomodation manager", which are not spesific to quick-service restaurant operations. Subsequently, the model considers and ultimately selects titles that emphasize leadership within this spesific restaurant context, narrowing down to "quick service restaurant team leader" and "fast food shift team leader" as the most apt job titles. The reasoning of the model is correct on chosing these titles for their precise reflection of the managerial and leadership responsibilities pertinent to the restaurant environment.

B. Job postings and Predicted ESCO job titles

The following tables provide examples of job titles, job posting descriptions, and the corresponding gold labels in Table 5 and optimized LLama-3 job titles in Table 6. These examples illustrate how the job titles assigned by recruiters may not always capture the specific nature of the job described in the postings. The gold labels and the optimized LLama-3 job titles offer a more accurate representation of the job roles based on the detailed job descriptions.

The job title "Commessa" (Salesperson) is generic and does not specify the specialization required for the job. The gold label "telecommunications equipment specialised seller" fits better because the job description clearly focuses on selling telecommunications equipment, which requires specific knowledge and skills related to this type of product. The gold label accurately reflects the specialized nature of the role. The job title "Project engineer" given by the recruiter suggests a technical and

Gold Label Job Title

Quick Service Restaurant Team Leader

Posting Job Title

Encargado de Franquicias

Posting Description:

- Responsable de garantizar la satisfacción de los huéspedes y de gestionar y superar los objetivos financieros y operativos de los restaurantes a mi cargo.
- Garantizar una excelente atención a los huéspedes en base a las promesas y estándares definidos.
- Liderar, motivar y desarrollar equipos.
- Facilitar los recursos y el apoyo necesario a los equipos en sus restaurantes.
- Utilizar de manera eficaz los diferentes recursos de la Compañía.
- Identificar oportunidades y amenazas de negocio en el mercado.
- Aportar ideas y ejecutando proyectos en el corto y medio plazo.
- Difundir las mejores practicas y resolver problemas comunes en los restaurantes.
- Cumplir los protocolos y políticas de la Marca y la Compañía.
- Garantizar y difundir los valores y principios definidos por la Compañía.

Skills: SAP Girnet Gtock, Cuiner

ESCO Job Titles:

Restaurant Manager, Business Manager, Hospitality Revenue Manager, Accommodation Manager, Delicatessen Shop Manager, Rooms Division Manager, Customer Experience Manager, Quick Service Restaurant Team Leader, Destination Manager, Membership Manager

Table 4
Spanish job posting Example

Posting Job Title	Job Posting Description	Gold Labels
Commessa	Commessa; Commessa; - Presentazione e vendita di attrezzature per telecomunicazioni ai clienti; - Servizio e supporto clienti; - Gestione delle transazioni di vendita; - Gestione dello stock e dell'inventario.	Telecommunications equipment specialised seller
Project Engineer	Project Engineer; Project Engineer; PROJECT MANAGER / PROJECT ENGINEER Divisione: Amministrazione Tecnica - Coordinamento delle attività di gestione progetti in ambito tecnico; - Supporto al Product Development; - Pianificazione e monitoraggio delle attività progettuali; - Supervisione del team tecnico; - Assistenza alla gestione dei fornitori e del budget di progetto.	Project manager, Product development manager

Table 5
Examples of Job Titles, Descriptions, and Gold Labels

engineering-focused role. However, the job description emphasizes project management, coordination of project activities, support to product development, and supervision of the technical team. The gold label “project manager” fits better as it captures the overall management and coordination responsibilities described, which are more aligned with the duties of a project manager than just a project engineer.

The job title "Addetto alle vendite" (Sales Assistant) is too generic and does not capture the specialized nature of the role described in the vacancy. The description specifies duties typical of a deli worker, such as serving customers, slicing cheeses and cured meats, preparing packages, and managing the deli counter. Our model’s titles “meat and meat products specialised seller” and “deli worker” are more precise, indicating a specialized role in food handling and customer service, which goes beyond the general sales assistant title. This demonstrates our model’s ability to interpret the specific context and responsibilities of the job accurately.

The job title “IT Specialist” is generic and could encompass various IT roles. However, the job description clearly indicates responsibilities such as managing ICT projects, coordinating a software development team, planning and monitoring development activities, managing ICT resources and budget, and providing advanced techni-

Posting Job Title	Job Posting Description	Optimized LLama-3 Job Titles
Addetto alle vendite	Addetto alle vendite; Addetto alle vendite; Salumiere: servizio clientela, tagli di formaggi e salumi, preparazione confezioni, gestione banco gastronomia.	Meat and meat products specialised seller, Deli worker, Food and beverage server
IT Specialist	IT Specialist; IT Specialist; Responsabile della gestione dei progetti ICT; Coordinamento del team di sviluppo software; Pianificazione e monitoraggio delle attività di sviluppo; Gestione delle risorse ICT e del budget; Assistenza tecnica avanzata e risoluzione dei problemi.	ICT project manager, Software development manager
Sales Manager	Sales Manager; Sales Manager; Sviluppo del business aziendale; Definizione delle strategie di vendita; Gestione del team di vendita; Monitoraggio delle performance e raggiungimento degli obiettivi di vendita; Gestione delle relazioni con i clienti chiave e i partner strategici.	Business development manager, Sales director

Table 6
Examples of Job Titles, Descriptions, and Optimized Job Titles

Table 7
Examples of Job Postings with Ambiguous Classification due to Multilingual and Contextual Challenges

Job Title	Description Excerpt	Labels Suggested
Junior Project Manager	Applicare i metodi e gli strumenti propri del Project Management a commesse specifiche per il settore dell'automazione industriale, di cui l'azienda fornisce sistemi di visione artificiale.	Project Manager, ICT Project Manager, Programme Manager
Assistente Amministrativo (Healthcare)	Gestione dei flussi delle segnalazioni dei cittadini per prenotazioni vaccinazioni e assistenza pandemica, inclusa la verifica del "certificato verde" per la conformità alle normative sanitarie.	Healthcare Assistant, Administrative Assistant, Contact Tracing Agent
Commesso di Negozio (Retail)	Creazione di vetrine accattivanti con abbinamenti di tendenza e assistenza alla clientela nella scelta dei prodotti.	Shop Assistant, Sales Assistant, Visual Merchandiser
Team Leader (Energy Sector)	Predisposizione documenti formativi e aggiornamento processi operativi presso sede Enel, inclusa l'implementazione e il collaudo di software per la gestione energetica.	Team Leader, Energy Analyst, Business Process Analyst
Assistente Amministrativo (Legal and Fiscal)	Compiti legati al Registro Nazionale delle Varietà Vegetali e mansioni fiscali complesse come Dichiarazioni IRAP.	Accounting Assistant, Administrative Assistant, Compliance Officer

cal support. The optimized titles “ICT project manager” and “software development manager” are more accurate as they reflect the leadership, coordination, and project management aspects of the role, which go beyond the scope of a general IT specialist.

The job title “Sales Manager” suggests a mid-level management role. However, the job description highlights responsibilities such as business development, defining sales strategies, managing the sales team, monitoring performance, and managing relationships with key clients and strategic partners. These responsibilities are more aligned with a higher-level role such as “business development manager” or “sales director”, which involve strategic planning and high-level management.

C. Ambiguity from Specialized and Contextual Factors

To further understand the complexity of job classification in a multilingual context, we conducted an ablation study focusing on cases where both human annotators and LLMs demonstrated shared uncertainty in assigning definitive labels. These cases were particularly challenging due to specialized terminology, regional language variations, or overlapping responsibilities within job postings. Table 7 highlights key examples where annotators, despite their recruitment expertise, aligned with the LLMs in experiencing ambiguity.

As presented in Table 7, each example illustrates specific challenges encountered in classifying job postings across multilingual and sector-specific contexts. The *Junior Project Manager* job posting, for instance, combines general project management with specialized tasks such as machine vision, but without enough specific context, it is unclear whether the focus should be on technical expertise or managerial skills. The *Project Engineer* example shows the impact of technical terminology and sector-spesific language on classification. Terms such as “SCADA” and “Modbus TCP” are common in international engineering contexts but may not align with typical understanding of recruiters, leading to the selection of varied labels by both LLMs and annotators. The example of the *Assistente Amministrativo* with a legal and fiscal focus involves highly specialized processes such as “Registro Nazionale delle Varietà Vegetali” and complex fiscal duties like “Dichiarazioni IRAP.” These terms relate to specific Italian government and regulatory compliance, which could exceed the annotators’ typical recruitment experience, thus resulting in generalized labels that do not fully capture the compliance and accounting complexity.

These cases emphasize that job postings, as human-created documents, often do not provide enough context for a definitive classification, resulting in ambiguity across specialized and regional terms.

D. Analysis of Model Alignment with Partial Agreement Ground Truth Labels

Table 8
Performance Metrics for Top 5 and Top 10 Predictions

Model	Precision		Recall	
	@5	@10	@5	@10
Spanish (SPA)				
llama-3-8b (CoT opt.)	0.12	0.06	0.58	0.62
llama-3-8b (CoT)	0.22	0.16	0.64	0.68
llama-3-8b (SkillGPT)	0.19	0.12	0.36	0.62
mBart-large-mnli (0-shot)	0.15	0.14	0.39	0.70
multilingual-e5-large	0.20	0.19	0.48	0.92
Italian (ITA)				
llama-3-8b (CoT opt.)	0.12	0.06	0.56	0.60
llama-3-8b (CoT)	0.23	0.07	0.55	0.59
llama-3-8b (SkillGPT)	0.22	0.06	0.53	0.59
mBart-large-mnli (0-shot)	0.27	0.06	0.31	0.58
multilingual-e5-large	0.35	0.08	0.39	0.79

In our evaluation, we established two levels of ground truth labels: *gold* and *silver*. Gold labels represent unanimous agreement among all three annotators (GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet), validated by human experts. Silver labels indicate a strong majority consensus, assigned when any two annotators agree, even if the third disagrees.

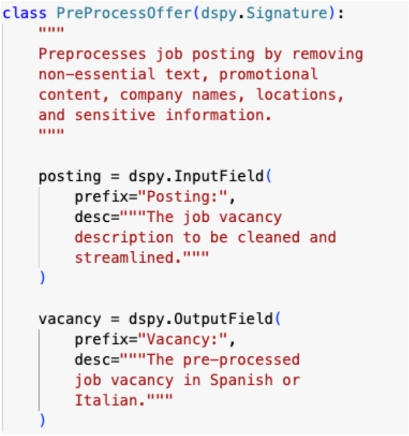
We assessed our model’s performance on both silver and gold labels to understand its effectiveness under different levels of agreement. We had reported results for gold labels in Table 2 and 3, results for silver label are presented in Table 8. For the Spanish dataset, the model’s performance was relatively consistent between silver and gold labels, with only minor variations in precision and recall. This consistency suggests that the model robustly captures underlying patterns in the job postings, regardless of labeling strictness.

In contrast, the Italian dataset exhibited more significant differences between performances on silver and gold labels. For example, in some cases, the precision was higher for silver labels while recall was higher for gold labels. This disparity may indicate that the model better captures broader classifications aligning with majority consensus in Italian but struggles with the stricter criteria required for unanimous agreement.

An interesting observation is that optimization using gold label ground truth data had a negative effect on the models’ scores derived from silver labels. This could be explained by the fact that during optimization, the language models became more attuned to the patterns present in the gold labels, potentially diverging from those in the silver labels. As a result, the models may have become less effective at predicting labels where only partial agreement (silver labels) was present among the automatic methods.

E. DSPy Signature

We utilize DSPy signatures to prompt large language models (LLMs) for performing downstream tasks. To optimize the script, recursive LLM calls were employed, resulting in its final form based on empirical observations.



```
class PreProcessOffer(dspy.Signature):
    """
    Preprocesses job posting by removing
    non-essential text, promotional
    content, company names, locations,
    and sensitive information.
    """

    posting = dspy.InputField(
        prefix="Posting:",
        desc="""The job vacancy
        description to be cleaned and
        streamlined."""
    )

    vacancy = dspy.OutputField(
        prefix="Vacancy:",
        desc="""The pre-processed
        job vacancy in Spanish or
        Italian."""
    )
```

Figure 4: Pre-processing Signature