Semantic Reranking for Zero-Shot Occupation-Skill Linking

System Description for the HR TalentCLEF 2026 Shared Task B

Alejandro Jesús Castañeira Rodríguez[1,*]

[1]JANZZ.technology, Zürich, Switzerland

Abstract

This paper presents our system description and methodology for the HR TalentCLEF 2026 Task B, focusing on the automated, semantic linking of labor market skills to standard occupations. Accurately mapping these relationships is crucial for decoding workforce dynamics, aligning educational curricula with rapidly shifting industry needs, and providing actionable upskilling suggestions to professionals navigating their career paths. To build a robust model capable of zero-shot inference on new occupations and skills and to mitigate the dependency on pre-existing lists of occupations associated with a skill, we designed a purely semantic matching pipeline. During data preparation, because official ESCO descriptions were not available at inference time, we generated comprehensive 400-word descriptions for occupations from scratch using the Gemma 4 Large Language Model. Conversely, we retained the original baseline ESCO skill descriptions to effectively disambiguate misleading skill titles. Our core architecture employs a computationally efficient two-stage retrieve-and-rerank pipeline relying on highly optimized, domain-adapted transformer models rather than computationally prohibitive massive LLMs. The first stage utilizes a dense bi-encoder based on `gte-modernbert-base` (149M parameters), fine-tuned with a contrastive, multitask learning objective on millions of parsed job descriptions. The second stage processes candidate pools using `Qwen3-Reranker` (600M parameters), fine-tuned on the official Task B dataset with Binary Cross Entropy (BCE) on graded relevance labels. Our multi-stage approach achieved highly competitive performance on the development set, yielding a General Relevance NDCG of 0.7844 and a Mean Reciprocal Rank (MRR) of 0.8831. Furthermore, the pipeline generalized effectively to the unseen official test set, achieving a Graded NDCG of 0.7399 and a Binary NDCG of 0.7746, proving its viability for scalable real-world labor market applications.

Keywords

Occupation-Skill Linking, Labor Market Dynamics, Career Pathways, HR TalentCLEF, Contrastive Learning, Information Retrieval, AI in Human Resources

1. Introduction

The global labor market is a highly dynamic ecosystem, characterized by the continuous emergence of new roles and the rapid, technology-driven evolution of required skill sets. Accurately identifying and linking the necessary skills to their corresponding occupations has become a critical necessity for multiple stakeholders. For policymakers and economists, these linkages decode broad workforce trends and mobility patterns. For educational institutions, understanding the precise capabilities demanded by the market is essential for closing the skills gap and aligning academic curricula with real-world industry needs. Most importantly, for individual professionals, granular occupation-skill mapping powers the next generation of career development platforms, enabling highly personalized skill suggestions, identifying transition pathways between roles, and guiding strategic upskilling.

This complex mapping challenge has increasingly been addressed by deep learning paradigms, building upon foundational transformer architectures [1] and robust pre-training methodologies [2] capable of capturing deep contextual meanings. The HR TalentCLEF 2026 shared task [3], specifically Task B [4], challenges researchers to effectively and automatically associate occupations with their required and contextual skills to facilitate these very applications.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

a.castaneira@janzz.technology (A. J. C. Rodríguez)

While structured frameworks are invaluable for providing standardized vocabularies, relying on pre-existing lists of occupations associated with a skill during model training can introduce dependencies that limit generalization. Specifically, fine-tuning a model by appending a known list of associated occupations directly to a skill creates a "chicken-and-egg" problem: the model learns to rely on this list to make accurate predictions, but for newly emerging skills, this list does not yet exist. Attempting to bypass this limitation by using Large Language Models (LLMs) to generate these associated lists can introduce critical errors and hallucinations that negatively bias the semantic space. Therefore, our approach seeks to pair occupations to skills based purely on the inherent semantic definitions of the concepts themselves, utilizing dense passage retrieval techniques [5] to ensure robust, zero-shot adaptability without relying on pre-existing structural connections.

2. Data Preparation and Context

The primary dataset utilized for our cross-encoder fine-tuning and overall evaluation was the official training data provided by the HR TalentCLEF 2026 Task B organizers [4]. This dataset provides curated mappings between standard occupation titles and their requisite skills (annotated as essential or optional), grounded in the ESCO (European Skills, Competences, Qualifications and Occupations) taxonomy [6]. While ESCO offers a highly structured and invaluable standard vocabulary, its textual descriptions are often too vague or brief to provide sufficient context for modern dense retrieval embeddings. To address this, we implemented an asymmetric augmentation strategy, expanding occupations while tightly constraining skills.

2.1. Occupation Description Generation

A significant challenge in applying pre-trained dense retrieval models to taxonomic tasks is the brevity and lack of detail in the target texts. Because official ESCO descriptions are not consistently available for occupations during inference time within the shared task, relying on sparse occupation titles alone leads to poor semantic overlap. Furthermore, we observed that many times the raw ESCO descriptions were simply not descriptive enough and repeatedly failed to consistently represent the underlying meaning of the occupation.

To counter this limitation, we decided to augment the structural representations by generating comprehensive 400-word descriptions from scratch using the instruction-tuned `google/gemma-4-E4B-it` large language model[1] [7]. This generation provided a much richer textual surface for the representation learning model, allowing it to anchor onto a wider array of semantic tokens and consistently align occupational semantics. Crucially, these generated descriptions were utilized to represent the occupations across the entire pipeline, replacing the sparse baselines during both model training and final inference. The prompt utilized for this task was carefully constrained to prevent the introduction of conversational artifacts:

```
prompt= Create an official description for the occupation: <occupation title> in no more
    than 400 words. Please, do not include your own comments/dialogue in the response:
```

This generation step allowed the model to grasp the nuanced daily responsibilities, required proficiencies, and contextual environment of each occupation, mimicking the rich, multi-faceted information necessary to make accurate career-pathing suggestions in real-world scenarios.

2.2. Skill Descriptions and Semantic Disambiguation

Conversely, we explicitly chose *not* to augment the skill definitions using a Large Language Model. Instead, we extracted and utilized the baseline, succinct ESCO descriptions for the skills.

We emphasize the strict inclusion of these original skill descriptions because ESCO titles and labels are frequently misleading or heavily context-dependent. The description of the skill often changes the

[1] https://huggingface.co/google/gemma-4-E4B-it

semantic space represented by the title entirely. If a semantic engine were to rely on augmented text or titles alone, the representation space would likely collapse due to over-generalization.

For example, the ESCO skill title *"identify available services"* is inherently ambiguous. Without its accompanying baseline description ("*Identify the different services available for an offender during probation in order to help in the rehabilitation and re-integration process, as well as advising the offenders as to how they can identify services available to them*"), a semantic model might incorrectly associate it with IT cloud service provisioning, retail customer service, or hospitality management. By relying exclusively on the baseline skill description, we enforce a strict semantic boundary, teaching the model to understand true requirements rather than training it on the "answers" via augmented lists.

3. Methodology

Our methodology utilizes a state-of-the-art two-stage Information Retrieval approach: a heavily optimized dense bi-encoder for rapid initial candidate retrieval, followed by a deeper cross-encoder [8] for fine-grained reranking.

3.1. Stage 1: Contrastive Bi-encoder Fine-Tuning

To establish a strong initial retrieval mechanism capable of zero-shot domain mapping, we trained a bi-encoder based on the `Alibaba-NLP/gte-modernbert-base` [9, 10, 11] architecture, which extends the foundational principles of bidirectional encoder representations [12]. To adapt the model to the specific semantic topology of the labor market, we constructed a domain-specific fine-tuning corpus. We parsed several million real-world job descriptions using the JANZZ.technology semantic parser [13]. From these raw postings, we extracted structured pairings of occupations and their explicitly required skills.

Figure 1 illustrates the architecture of our bi-encoder training phase. Using this parsed corpus, we fine-tuned the model with a contrastive learning objective [14], employing a multitask format to maximize generalization. The training data was constructed into two complementary pair formulations:

1. **Fill-in-the-Blank Formulation**: We concatenated an occupation with all its parsed skills, intentionally leaving out one target skill. The dataset pair became (`Occupation +` $[Skill_1, \ldots, Skill_{N-1}]$, `Target` $Skill_N$). This objective forces the model to semantically predict what capability is missing from the provided occupational context, effectively learning functional dependencies.
2. **Context Prediction Formulation**: To further robustify the embeddings, we inverted the relationship: (`Target` $Skill_N$, `Occupation +` $[Skill_1, \ldots, Skill_{N-1}]$). The model must predict the surrounding professional context given a standalone skill. This is conceptually analogous to the continuous skip-gram method utilized in Word2Vec [15], extended here to dense sentence embeddings.

We fine-tuned the model using the `SentenceTransformers` library [16, 17]. Because the sheer number of combinations between skills and professional contexts grew exponentially, we utilized the `CachedGIST` loss function. This allowed us to circumvent the problem of heavily penalizing highly similar skill-context combinations as false negatives within a batch, while simultaneously allowing for a significantly larger batch size necessary for effective contrastive learning. All experiments were conducted on a single NVIDIA RTX 5090 GPU. Given the vast permutation space of occupation-skill pairings, computational constraints limited the training duration to one-quarter (0.25) of a single epoch, which required approximately two full days of continuous compute.

3.2. Stage 2: Hard Negative Mining and Reranking

While bi-encoders are highly efficient for retrieval, they lack the capacity to model deep token-level interactions between the query and the document. For the secondary reranking stage, we utilized the

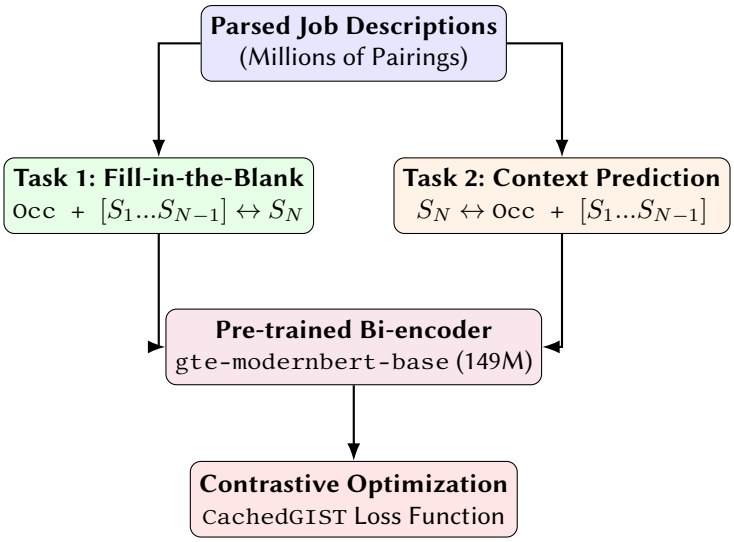

Figure 1: Visual representation of the Stage 1 contrastive fine-tuning pipeline, detailing the dual formulations extracted from job descriptions.

official HR TalentCLEF Task B training data to train a cross-encoder. We built a dense index of all skills in the training dataset using our fine-tuned bi-encoder. For each occupation, we retrieved k hard negative skills (where k equals the total amount of positive and optional skills for the given title in the official labels). This hard negative mining strategy is crucial for forcing the reranker to distinguish between semantically related but functionally incorrect skills.

We then utilized `Qwen/Qwen3-Reranker-0.6B` [18, 19] as our cross-encoder architecture. The model was fine-tuned on the TalentCLEF dataset using the `SentenceTransformers` library [16, 17] with a Binary Cross-Entropy (BCE) loss function. We employed soft labels based on the official skill criticality annotations: **1.0** for essential skills, **0.65** for optional skills, and **0.10** for mined unrelated skills. At inference time, the fine-tuned `Qwen3-Reranker` is used to evaluate the cross-attention between an occupation and candidate skills, ultimately scoring and ranking the top 1,000 skills for each occupation title in both the development and test sets.

3.3. Computational Efficiency and Scalability

A fundamental design principle of our approach is computational viability. In the current landscape of Natural Language Processing (NLP), there is a pervasive trend toward utilizing massive, billion-parameter Large Language Models (LLMs) for retrieval and classification tasks. While these models offer high baseline performance, their immense memory footprint and severe inference latency make them highly impractical for real-time, scalable applications within production HR platforms or educational matching engines.

By intentionally architecting our system around smaller, highly optimized models, we achieve domain adaptability without prohibitive computational overhead. The first-stage bi-encoder, based on `gte-modernbert-base`, utilizes only 149 million parameters, allowing for ultra-fast dense index querying. Our second-stage cross-encoder, `Qwen3-Reranker-0.6B`, operates efficiently at 600 million parameters. This specific pairing demonstrates that rigorous contrastive fine-tuning on domain-specific corpora, coupled with hard negative mining, allows smaller models to deeply understand complex semantic spaces. The resulting pipeline guarantees rapid inference and sustainable deployment costs, serving as a highly effective alternative to relying on massive, generalized LLMs for highly specific labor market tasks.

4. Results

Our approach was evaluated on both the Task B development set and the unseen official test set using standard Information Retrieval metrics. The evaluation considers General (Binary) Relevance and Graded Relevance.

To explicitly isolate the performance contributions of each design choice, we first present the raw performance scores across all models in Table 1.

Table 1
Retrieval Performance Comparison on the Development Set

	General Relevance				Graded Relevance			
Metric	**Base Bi-enc.**	**FT Bi-enc.**	**+ Base Qwen3**	**+ FT Qwen3**	**Base Bi-enc.**	**FT Bi-enc.**	**+ Base Qwen3**	**+ FT Qwen3**
NDCG	0.7244	0.7514	0.7655	**0.7844**	0.6891	0.7059	0.7352	**0.7527**
MAP	0.2177	0.2682	0.2893	**0.3272**	0.2177	0.2682	0.2893	**0.3272**
MRR	0.7804	0.8002	0.8401	**0.8831**	0.7804	0.8002	0.8401	**0.8831**
Precision@5	0.5803	0.6059	0.6382	**0.6882**	0.5803	0.6059	0.6382	**0.6882**
Precision@10	0.5372	0.5826	0.5872	**0.6414**	0.5372	0.5826	0.5872	**0.6414**
Precision@100	0.3378	0.3857	0.4134	**0.4603**	0.3378	0.3857	0.4134	**0.4603**

To further highlight the value of our methodology, Table 2 isolates the cascading marginal improvements (Δ) introduced by each consecutive stage relative to the baseline bi-encoder.

Table 2
Cascading Component Improvements (Δ) relative to the initial bi-encoder

	General Relevance Δ Gains			Graded Relevance Δ Gains		
Metric	**+ FT Bi-encoder**	**+ Base Qwen3 Reranker**	**+ FT Qwen3 Reranker**	**+ FT Bi-encoder**	**+ Base Qwen3 Reranker**	**+ FT Qwen3 Reranker**
Δ NDCG	+0.0270	+0.0411	**+0.0600**	+0.0168	+0.0461	**+0.0636**
Δ MAP	+0.0505	+0.0716	**+0.1095**	+0.0505	+0.0716	**+0.1095**
Δ MRR	+0.0198	+0.0597	**+0.1027**	+0.0198	+0.0597	**+0.1027**
Δ Precision@5	+0.0256	+0.0579	**+0.1079**	+0.0256	+0.0579	**+0.1079**
Δ Precision@10	+0.0454	+0.0500	**+0.1042**	+0.0454	+0.0500	**+0.1042**
Δ Precision@100	+0.0479	+0.0756	**+0.1225**	+0.0479	+0.0756	**+0.1225**

The cascading effect clearly illuminates how each architectural layer builds upon the last. Domain-specific contrastive training initializes a sturdy representation framework, directly adding substantial baseline improvements (e.g., **+0.0505 MAP**). Laying the un-tuned cross-encoder token interaction layer over this foundation offers a secondary compounding lift, pushing **MRR** up by **+0.0597**. Finally, fine-tuning the cross-encoder with targeted, soft-labeled criticality constraints delivers the most pronounced cascading impact, capturing nuanced mappings to secure ultimate gains of **+0.1027 MRR** and **+0.1095 MAP** over the original bi-encoder.

Test Set Performance: To validate the generalization of our semantic approach, the final two-stage pipeline was evaluated on the official unseen test set. The model maintained highly competitive performance, achieving a **Graded NDCG of 0.7399** and a **Binary (General) NDCG of 0.7746**. Furthermore, the system achieved a **MAP of 0.3093** and an **MRR of 0.8052**. This demonstrates that our zero-shot focused methodology successfully maps unseen conceptual definitions without overfitting to the predefined training lists.

5. Conclusion

Accurate occupation-skill linking serves as the foundational layer for modern workforce analytics, educational planning, and personalized career development. In this paper, we outlined our methodology for HR TalentCLEF 2026 Task B, designing a system aimed at solving these core labor market challenges efficiently. We demonstrated that generating occupation descriptions with LLMs from scratch, while rigidly anchoring skill semantics to baseline ESCO definitions, successfully prevents representation collapse and enables true semantic matching.

Furthermore, our experiments revealed powerful architectural alignment when stacking computationally efficient semantic encoders: fine-tuning a dense 149M parameter bi-encoder on millions of parsed job descriptions sets a highly robust foundation, which the 600M parameter `Qwen3-Reranker` capitalizes on to achieve highly competitive results on the development set. Crucially, our performance generalized strongly to the unseen official test set (Binary NDCG: 0.7746), validating the viability of our purely semantic, list-independent matching pipeline to adapt to the ever-evolving demands of the global labor market without requiring prohibitive infrastructural scale.

Acknowledgments

We thank the HR TalentCLEF 2026 organizers for providing the dataset and evaluation framework that made this research possible.

Declaration on Generative AI

During the preparation of this work, the author used the Gemma 4 Large Language Model in order to generate comprehensive textual descriptions from scratch for occupations in the training dataset to provide a richer semantic surface for model training. Additionally, generative AI was utilized to assist in writing, editing, proofreading, and improving the manuscript text itself. After using these tools, the author reviewed and edited all content as needed and takes full responsibility for the publication's content and the resulting dataset's integrity.

References

[1] A. Vaswani, et al., Attention is all you need, in: Advances in neural information processing systems, 2017.
[2] Y. Liu, et al., RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[3] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, W. Veys, C. P. Carrino, M. De Lange, D. Deniz Cerpa, A. Rodrigo, J.-J. Decorte, R. Zbib, Overview of the talentclef 2026: Skill and job title intelligence for human capital management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2026.
[4] W. Veys, L. Gasco, M. De Lange, J.-J. Decorte, Overview of talentclef 2026: Task b — job-skill matching with skill type classification, in: CLEF 2026 Working Notes, 2026.
[5] V. Karpukhin, et al., Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 EMNLP, 2020.
[6] M. Zhang, R. van der Goot, B. Plank, ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.
[7] Gemma Team, Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[8] R. Nogueira, K. Cho, Passage re-ranking with bert, in: arXiv preprint arXiv:1901.04085, 2019.
[9] Z. Li, et al., Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).
[10] X. Zhang, et al., mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, in: Proceedings of the 2024 EMNLP Industry Track, 2024, pp. 1393–1412.
[11] W. Antoun, B. Sagot, D. Seddah, ModernBERT or DeBERTaV3? Examining Architecture and Data Influence on Transformer Encoder Models Performance, arXiv preprint arXiv:2504.08716 (2025).
[12] J. Devlin, et al., BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 NAACL-HLT, 2019.

[13] JANZZ.technology, JANZZ.technology: Multilingual ontology and semantic parsing for the global labor market, https://janzz.technology/, 2026.
[14] T. Gao, et al., SimCSE: Simple contrastive learning of sentence embeddings, arXiv preprint arXiv:2104.08821 (2021).
[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013.
[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, C. Paul, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.
[18] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025).
[19] M. Li, et al., Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, arXiv preprint arXiv:2601.04720 (2026).