ui-nlp at TalentCLEF 2026: Multilingual Bi-Encoder Retrieval with Ensemble Strategy for Job-Person and Job-Skill Matching

Notebook for the TalentCLEF Lab at CLEF 2026

Ibtihal Qomariyyah Luthfiyyah[1,*], Satrio Yudhoatmojo[1] and Indra Budi[1]

[1]*Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia*

Abstract

This paper describes the system submitted by ui-nlp to TalentCLEF 2026, participating in two tasks: Task A (Contextualized Job-Person Matching) and Task B (Job-Skill Matching with Skill Type Classification). The proposed system employs a multilingual bi-encoder retrieval framework based on `intfloat/multilingual-e5-large-instruct`, combined with task-specific instruction prefixes and a score-based ensemble strategy. For Task A, job descriptions and candidate resumes are encoded independently into a shared semantic space, enabling efficient multilingual retrieval on the English and Spanish test sets. For Task B, each skill from the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy is expanded using all available aliases to enrich semantic coverage, and the model is further adapted through fine-tuning on essential job-skill pairs from the official training data. The experimental results show that the proposed system consistently outperforms the official TalentCLEF 2026 baselines on both tasks. For Task A, relative improvements of 58.1%, 69.7%, and 68.3% in Mean Average Precision (MAP) are achieved in the English monolingual (en-en), Spanish monolingual (es-es), and Cross-Lingual (en-es) test sets. For Task B, the ensemble achieves relative improvements of 9.7% in Normalized Discounted Cumulative Gain (nDCG) binary, 9.9% in nDCG graded, and 60.8% in MAP on the test set.

Keywords

multilingual retrieval, bi-encoder, dense retrieval, job-person matching, job-skill matching, TalentCLEF, ensemble strategy, sentence embeddings, ESCO taxonomy

1. Introduction

TalentCLEF 2026 is a multilingual benchmark that focuses on the application of Natural Language Processing (NLP) to support Human Capital Management (HCM) processes, particularly tasks related to matching jobs, skills, and candidates in multilingual scenarios [1]. In this study, we participated in two main tasks: Task A: Contextualized Job-Person Matching and Task B: Job-Skill Matching with Skill Type Classification.

Task A in TalentCLEF 2026 focuses on retrieving and ranking the most relevant candidates for a given job description (JD). Unlike Task A in TalentCLEF 2025, which focused on multilingual job title matching where the system only needed to match short textual representations of job titles with semantically similar job titles [2], Task A 2026 involves more complex contextual information because both job descriptions and candidate resumes contain lengthy and heterogeneous content, including work experience, skills, education, certifications, and job responsibilities [3]. Furthermore, Task A includes multilingual and cross-lingual retrieval scenarios, requiring systems to effectively capture semantic relationships across languages [3].

Meanwhile, Task B in TalentCLEF 2026 focuses on job-skill matching while considering skill type classification. In contrast to Task B in TalentCLEF 2025, which addressed job title-based skill prediction by predicting relevant skills solely from short job title representations [2], Task B 2026 requires a deeper understanding of the semantic relationships between occupations and skills while simultaneously distin-

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

*Corresponding author.

ibtihal.qomariyyah@ui.ac.id (I. Q. Luthfiyyah); satrio.baskoro@cs.ui.ac.id (S. Yudhoatmojo); indra@cs.ui.ac.id (I. Budi)

guishing different types of skills relevant to a particular job [4]. This task is challenged by high lexical variability, where a single skill may be expressed using different terms or synonyms. Consequently, keyword-matching approaches often struggle to capture the underlying semantic relationships between occupations and relevant skills [4].

2. Related Work

2.1. Job-Person Matching

Research on candidate-job matching has evolved from keyword-based and manually engineered feature approaches toward semantic representation-based methods. Ajjam et al. [5] proposed a semantic job matching framework based on Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity, demonstrating improved performance compared to traditional keyword-matching approaches. Although relying on relatively simple text representations, their study highlighted the importance of semantic information in improving the quality of candidate-job matching.

Subsequent developments have been dominated by transformer-based models for learning richer semantic representations. Xie and Tan [6] developed a Bidirectional Encoder Representations from Transformers (BERT)-based Siamese Network architecture that maps candidate profiles and job descriptions into a shared embedding space. Similarly, Rosenberger et al. [7] introduced CareerBERT, which leverages Sentence-BERT to generate unified vector representations of resumes and job descriptions. Furthermore, Myneni et al. [8] combined transformer-based representations with multiple matching signals, including skills, experience, and semantic similarity, to improve candidate ranking performance. Collectively, these studies demonstrate that embedding-based approaches have become the dominant paradigm in modern candidate-job matching systems.

2.2. Job-Skill Matching

As the demand for more accurate skill recommendation systems continues to grow, researchers have increasingly adopted semantic representations to model the relationship between occupations and skills. Alonso et al. [9] proposed a transformer-based approach combined with the O*NET taxonomy for job matching and skill recommendation. By leveraging sentence embeddings and semantic similarity, their framework was able to identify relevant skills beyond simple keyword overlap, demonstrating the effectiveness of semantic representations in HCM applications.

A similar direction was explored by Herold et al. [10], who addressed the task of matching competency frameworks with job advertisements. Their approach combined Natural Language Processing techniques and BERTopic to identify and map competencies relevant to labor market demands. The results showed that semantic representation-based methods improve the alignment between occupations and competencies compared to rule-based or lexical matching approaches. Nevertheless, most existing studies have focused on skill recommendation or competency mapping in monolingual settings, leaving multilingual job-skill matching scenarios relatively underexplored.

2.3. Multilingual Embedding Retrieval

Several studies in the TalentCLEF 2025 benchmark have demonstrated the effectiveness of multilingual dense retrieval approaches. Nizami et al. [11] proposed a retrieval framework based on Sentence-BERT and MPNet embeddings, combined with curriculum fine-tuning and hybrid retrieval strategies to improve multilingual job title matching performance. Similarly, Villar et al. [12] explored the use of modern embedding models and reranking techniques to enhance retrieval quality for the same task. More recent work has also investigated multilingual-E5 models, which are specifically designed for semantic retrieval and have shown strong performance in multilingual settings [13, 14].

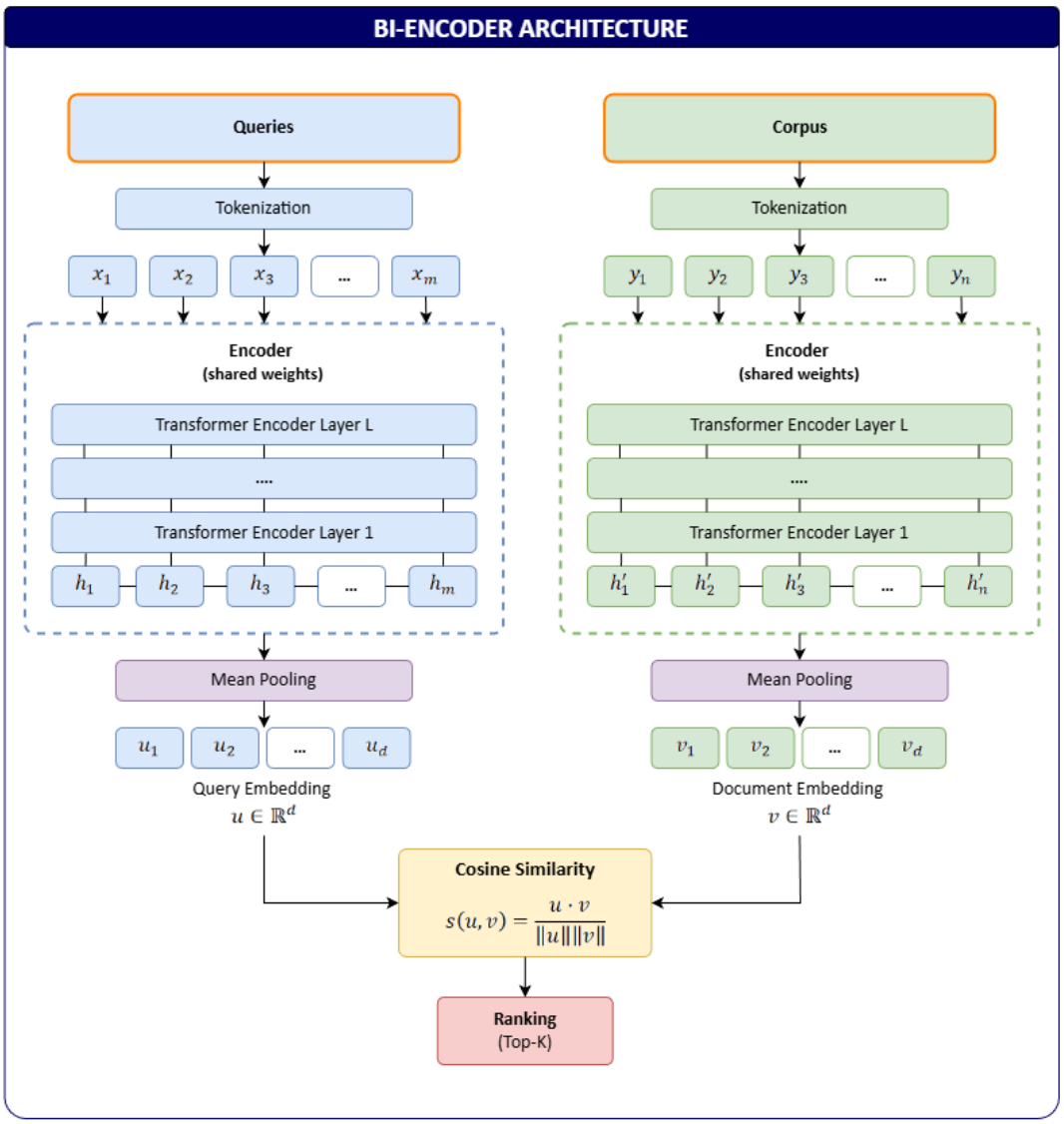

Figure 1: Overview of the proposed multilingual bi-encoder retrieval framework. Job descriptions or job titles are encoded as queries, while resumes or skills are encoded as the corpus in a shared semantic space.

3. System Description

3.1. Bi-Encoder Retrieval Framework

The proposed system employs a bi-encoder architecture as the primary retrieval framework for both TalentCLEF 2026 tasks. An overview of the framework is presented in Figure 1.

Figure 1 illustrates the bi-encoder retrieval framework used in both Task A and Task B. Queries and corpus (orange-bordered boxes) represent inputs that differ from TalentCLEF 2025 approaches, encompassing richer contextual information beyond short job titles. Both inputs are independently tokenized and encoded using a shared transformer encoder. The resulting token-level representations are aggregated through mean pooling to produce fixed-dimensional dense embeddings $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Cosine similarity between the query and corpus embeddings is then computed to rank the corpus (Top-K).

Given a query q and a corpus c, their embeddings are computed as

$$u = f(q), \qquad v = f(c) \tag{1}$$

where $f(\cdot)$ denotes the shared encoder. The relevance score is then calculated using cosine similarity:

$$s(u, v) = \frac{u \cdot v}{\|u\|\|v\|} \tag{2}$$

In Task A, the query consists of a job description (JD) and the corpus corresponds to a candidate resume. Both text types are typically long and contain diverse information, making rich semantic representations essential for capturing the relevance between them.

In Task B, the query is a relatively short job title, while the corpus consists of skills from the ESCO (European Skills, Competences, Qualifications and Occupations) taxonomy along with all their aliases. Each skill is represented by expanding (*exploding*) all available aliases to enrich the semantic coverage of skill candidates. Deduplication is then applied to the final results to ensure each skill appears only once in the ranking.

The bi-encoder is preferred over the cross-encoder due to its higher computational efficiency for large-scale retrieval. Unlike cross-encoders, which perform full interaction over each query-document pair, bi-encoders produce dense embeddings independently, allowing document embeddings to be precomputed and stored for efficient retrieval [15, 16].

3.2. Model Selection

We select `intfloat/multilingual-e5-large-instruct` (mE5-large-instruct) as the primary model for both tasks. This model is an XLM-RoBERTa-large-based encoder with 560 million parameters, trained using contrastive learning on large-scale multilingual data [17]. The model supports over 100 languages, including English and Spanish used in TalentCLEF 2026, enabling a single model to handle all language pairs (en-en, es-es, en-es) without requiring language-specific models. Furthermore, it supports instruction prefixes that allow queries to be encoded with task-specific context. The use of instruction prefixes has been shown to improve retrieval performance by enabling the model to align its vector representations with the specific objective of each task [18]. The instruction prefixes used for each task are described in detail in Sections 3.4 and 3.5.

In addition to mE5-large-instruct, several models are evaluated as baselines and comparisons. `paraphrase-multilingual-MiniLM-L12-v2` and `all-MiniLM-L6-v2` serve as the official baselines provided by TalentCLEF 2026 organizers for Task A and Task B, respectively. `BAAI/bge-large-en-v1.5` is included to assess how well a strong monolingual English retrieval model competes against multilingual ones [19]. `intfloat/multilingual-e5-large` is evaluated to measure zero-shot multilingual retrieval performance without instruction prefix [17]. For Task A, a hybrid BM25 combined with mE5-large is explored to assess whether sparse retrieval complements dense retrieval, and a cross-encoder reranker (`cross-encoder/ms-marco-MiniLM-L-12-v2`) is evaluated as an alternative, following the standard retrieve-and-rerank pipeline: the top-50 candidates retrieved by mE5-large-instruct are re-scored by the cross-encoder, since cross-encoders jointly encode each query-document pair and therefore cannot benefit from precomputed embeddings, making full-corpus reranking impractical at scale. For Task B, a hybrid semantic lookup combining dense retrieval with knowledge graph information from the official occupation-to-skill mapping (labeled as essential or optional).

3.3. Ensemble Strategy

To further improve retrieval performance, we employ a score-based ensemble strategy that combines the outputs of two retrieval models. The ensemble is motivated by the observation that different models may capture complementary semantic signals. Consequently, combining their retrieval scores can potentially produce more robust rankings than relying on a single model.

Since the score distributions produced by different models may vary substantially, the retrieval scores are first normalized using min-max normalization on a per-query basis:

$$s_i^{\text{norm}} = \frac{s_i - s_{\min}}{s_{\max} - s_{\min}} \tag{3}$$

where s_i denotes the score of document i, while $s_{\min}$ and $s_{\max}$ represent the minimum and maximum retrieval scores for a given query, respectively.

The final ensemble score is computed as a weighted linear combination of the normalized scores:

$$s_{\text{ens}} = \alpha s_1^{\text{norm}} + (1 - \alpha) s_2^{\text{norm}} \tag{4}$$

where α controls the contribution of the first model and $(1 - \alpha)$ represents the contribution of the second model.

The optimal ensemble weight was selected through grid search on the development set using candidate values $\alpha \in \{0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.9\}$. MAP was used as the optimization metric for Task A, whereas nDCG was used for Task B.

For Task A, the ensemble combines mE5-large-instruct and mE5-large. These two models were selected as ensemble candidates because they were the strongest-performing individual models identified in our zero-shot evaluation (Table 4). Weight optimization was performed independently for each language pair to account for differences in multilingual retrieval behavior. Although mE5-large and mE5-large-instruct share the same XLM-RoBERTa-large backbone, differing primarily in instruction tuning, they produce different representations due to their asymmetric encoding strategy: mE5-large-instruct encodes queries with task-specific instruction prefixes while encoding documents without, whereas mE5-large applies symmetric encoding to both.

Ensembling with an architecturally distinct model was empirically explored by combining BGE-large-en-v1.5 with both mE5-large and mE5-large-instruct. The mE5-large + BGE-large-en-v1.5 combination underperformed the proposed ensemble on both languages (Table 4). The mE5-large-instruct + BGE-large-en-v1.5 combination achieved a slightly higher MAP than the proposed ensemble on English but a lower MAP on Spanish, resulting in a lower average MAP overall. Both patterns are consistent with BGE-large-en-v1.5's weak monolingual Spanish performance observed earlier, which limits its suitability as an ensemble component for multilingual settings despite competitive English performance. We therefore retained mE5-large-instruct and mE5-large as the final ensemble components, as they provided more balanced gains across both languages, though further exploration of architecturally diverse combinations remains a direction for future work.

For Task B, the ensemble combines the zero-shot and fine-tuned versions of mE5-large-instruct. The zero-shot model provides strong multilingual generalization, whereas the fine-tuned model incorporates additional domain-specific knowledge obtained from the official occupation-to-skill training data.

3.4. Task A: Contextualized Job-Person Matching

For Task A, each job description (JD) is treated as a query, while candidate resumes constitute the retrieval corpus. Both job descriptions and resumes are encoded independently using the multilingual bi-encoder described in Section 3.1. The resulting dense embeddings enable the system to capture semantic relationships between job requirements and candidate qualifications, even when different terminology is used.

For instruction-based retrieval, each job description is formatted using the following query template. This instruction template was kept in English across all language settings (en-en, es-es, and en-es). The prefix was not translated into Spanish.

> Instruct: Given a job description, retrieve the most relevant resume
> Query: {job description}

After encoding, cosine similarity is computed between the query embedding and all resume embeddings. Candidates are then ranked according to their similarity scores, and the resulting ranking is used as the final retrieval output.

Task A is evaluated under three official benchmark settings: Overall Multilingual, Cross-Lingual, and Bias-Controlled. The Overall Multilingual setting evaluates general multilingual retrieval performance using Mean Average Precision (MAP) as the primary metric, whereas the Cross-Lingual setting specifically measures the ability to retrieve candidates across different languages, also using MAP as the primary metric. The Bias-Controlled setting evaluates retrieval robustness under benchmark conditions designed to reduce potential demographic biases, using Rank Biased Overlap (RBO) as the primary metric. For the Cross-Lingual (en-es) and Bias-Controlled settings, local evaluation was not possible as the required development data for these settings were not publicly available.

3.5. Task B: Job-Skill Matching with Skill Type Classification

For Task B, job titles are used as queries, while skills from the ESCO taxonomy constitute the retrieval corpus. To enrich semantic coverage, each skill is expanded using all available aliases provided in ESCO. This strategy allows the retrieval model to capture a wider range of skill expressions and terminology variations.

For instruction-based retrieval, each job title is formatted using the following query template. As in Task A, this instruction prefix was used in English without translation.

> Instruct: Given a job title, retrieve the most relevant skills for this job
> Query: {job title}

All expanded skill representations are encoded using the multilingual bi-encoder framework. Cosine similarity is then computed between the job title embedding and all candidate skill embeddings. Since multiple aliases may correspond to the same ESCO skill, a deduplication step is applied after retrieval to ensure that each skill appears only once in the final ranking.

Task B is evaluated under two official benchmark settings: Binary Relevance and Graded Relevance. In the Binary Relevance setting, skills are considered either relevant or non-relevant. In contrast, the Graded Relevance setting differentiates between core skills and contextual skills, providing a more fine-grained assessment of ranking quality.

4. Experimental Setup

4.1. Dataset

This study uses the official TalentCLEF 2026 datasets, which are publicly available through Zenodo [20]. The benchmark consists of two tasks with different retrieval objectives and data structures. Task A focuses on candidate-job matching using job descriptions and resumes, whereas Task B focuses on job-skill matching using occupations and skills derived from the ESCO taxonomy. Summary statistics for both datasets are provided in Tables 1 and 2.

4.1.1. Task A Dataset

Task A consists of job descriptions (queries) and candidate resumes (corpus) in English and Spanish. The dataset is provided in three evaluation settings: English monolingual (en-en), Spanish monolingual (es-es), and cross-lingual (en-es), where queries are written in English and resumes are written in Spanish.

Table 1 summarizes the statistics of the Task A dataset. No official training set is provided for this task. The development set includes relevance judgments (qrels) using binary relevance labels, where each job description is associated with exactly one relevant resume.

Table 1
Statistics of the Task A dataset.

Split	Language	Queries (JD)	Corpus (Resume)	Qrels
Development	English	10	472	472
Development	Spanish	10	472	472
Test	English	40	476	—
Test	Spanish	40	476	—

4.1.2. Task B Dataset

Task B consists of job titles as queries and ESCO skills as the retrieval corpus. Unlike Task A, an official training set is provided, containing occupation-skill pairs annotated as either essential or optional.

Table 2 presents the statistics of the Task B dataset. Relevance judgments follow a graded relevance scheme consisting of two levels: core skills (relevance score = 2) and contextual skills (relevance score = 1). Core skills are competencies considered essential and directly required for performing a given occupation, whereas contextual skills are those that are commonly associated with a role but not strictly mandatory. The development set contains 16,161 core skill pairs and 40,256 contextual skill pairs.

Table 2
Statistics of the Task B dataset.

Split	Queries (Job Title)	Corpus (Skills)	Qrels
Training	—	—	114,699
Development	304	9,052	56,417
Test	1,139	5,358	—

The training set was used exclusively for fine-tuning in Task B, while all model selection and hyperparameter tuning were performed on the development set.

4.2. Implementation Details

4.2.1. Hardware and Software

All experiments were conducted on a server equipped with an NVIDIA GeForce RTX 5060 Ti GPU with 16 GB of VRAM. The system was implemented in Python 3.12 using the Sentence-Transformers, Transformers (Hugging Face), and PyTorch libraries.

4.2.2. Zero-Shot Retrieval

The `intfloat/multilingual-e5-large-instruct` model was used directly without additional fine-tuning to generate dense semantic embeddings. Queries were encoded using task-specific instruction prefixes (see Sections 3.4 and 3.5), whereas documents were encoded without instruction prefixes. Encoding was performed with a batch size of 16 and a maximum sequence length of 512 tokens. Cosine similarity was then computed between query embeddings and document embeddings to produce the final ranking.

For Task A, the retrieval corpus consisted of candidate resumes, while job descriptions were treated as queries. For Task B, job titles were used as queries and ESCO skills as the retrieval corpus.

4.2.3. Fine-Tuning for Task B

For Task B, additional domain adaptation was performed by fine-tuning `intfloat/multilingual-e5-large-instruct` using the `SentenceTransformerTrainer` framework from the Sentence-Transformers library. Training data were constructed from 5,000

randomly sampled *essential* job-skill pairs extracted from the official occupation-to-skill mapping. Only *essential* pairs were used as positive examples because they provide stronger relevance signals than *optional* pairs. The model was optimized using `MultipleNegativesRankingLoss`, which is widely used for retrieval-oriented contrastive learning. Table 3 summarizes the training hyperparameters.

Table 3
Fine-tuning hyperparameters for Task B.

Hyperparameter	Value
Learning Rate	2e-6
Batch Size	16
Gradient Accumulation Steps	2
Epochs	1
Warmup Ratio	0.1
LR Scheduler	Cosine
Loss Function	MultipleNegativesRankingLoss
Optimizer	AdamW
Precision	float32

The learning rate and sample size were determined through small set of preliminary trials on the development set, evaluated using nDCG and MAP. Initial attempt using the full essential-pair set (62,480 pairs) with learning rates of 1×10^{-5} and 2×10^{-5}, trained via the legacy `fit()` interface, resulted in catastrophic forgetting. To recover stable training, we reduced the learning rate by an order of magnitude (2×10^{-6}) and switched to the `SentenceTransformerTrainer` API, which provides more stable optimization and supports gradient accumulation, beneficial when fine-tuning large multilingual models under GPU memory constraints. The smaller 5,000-pair sample was used at this stage to verify training stability before considering further scaling. Due to time constraints, we did not conduct a systematic grid search over learning rate or sample size, nor did we scale training to the full dataset. The effect of this configuration on retrieval performance is reported in Section 5.2.

5. Results and Analysis

5.1. Task A Results

Table 4 presents the evaluation results on the Task A development set using MAP as the primary evaluation metric. Development results are reported for the English (en-en) and Spanish (es-es) monolingual settings only. Results for the Cross-Lingual (en-es) and Bias-Controlled settings are not presented, as explained in Section 3.4.

Table 4
Comparison of retrieval methods on the Task A development set measured by MAP.

Method	English	Spanish	Average
paraphrase-multilingual-MiniLM-L12-v2 (baseline)	0.7060	0.6470	0.6765
BGE-large-en-v1.5	0.8482	0.5032	0.6757
mE5-large	0.8544	0.8774	0.8659
mE5-large + BGE-large-en-v1.5	0.8866	0.8629	0.8748
Hybrid BM25 + mE5-large	0.7737	0.7203	0.7470
Cross-encoder reranker top-50	0.5815	0.5741	0.5778
mE5-large-instruct	0.8989	0.8972	0.8980
mE5-large-instruct + BGE-large-en-v1.5	0.9012	0.8956	0.8984
Ensemble (proposed)	0.9000	0.9031	0.9015

Overall, multilingual dense retrieval models consistently outperformed the official TalentCLEF 2026 baseline. Among all individual models, mE5-large-instruct achieved the best performance, obtaining

a MAP of 0.8989 on English and 0.8972 on Spanish. Compared to the official baseline (paraphrase-multilingual-MiniLM-L12-v2), this corresponds to relative improvements of 27.3% and 38.7%, respectively. These results suggest that the combination of multilingual pretraining and instruction-based query encoding is beneficial for candidate-job matching.

In contrast, BGE-large-en-v1.5 exhibited a substantial performance gap between the two languages. While it achieved a strong MAP of 0.8482 on English, its performance dropped considerably to 0.5032 on Spanish. This finding highlights the limitations of monolingual retrieval models in multilingual settings, particularly when the target language is not well represented during pretraining.

It should be noted that the instruction prefix was kept in English across all settings (Section 3.4), meaning that for the Spanish monolingual setting, the query combines an English instruction with a Spanish dataset. Since the underlying XLM-RoBERTa-large backbone is pretrained on a multilingual corpus in which English is typically the most heavily represented language, this design choice could in principle favor English-language inputs. However, the close performance between English and Spanish suggests that this asymmetry did not noticeably disadvantage the Spanish setting, indicating that the model is able to align the semantic content of the job description with candidate resumes regardless of the instruction language.

The Hybrid BM25 + mE5-large approach and the cross-encoder reranker also produced lower performance than pure dense retrieval. Hybrid BM25 + mE5-large achieved MAP scores of 0.7737 and 0.7203 on English and Spanish, respectively, while the cross-encoder reranker obtained the lowest scores among all evaluated methods. One possible explanation is the presence of domain mismatch, as both BM25 and the MS MARCO-based reranker were originally designed for web search tasks, whereas Task A involves long and highly structured documents such as job descriptions and resumes.

The best overall performance was obtained by the proposed ensemble strategy, which combines mE5-large-instruct and mE5-large. Prior to score fusion, retrieval scores from both models were normalized using min-max normalization on a per-query basis. The optimal ensemble weights were determined independently for each language pair using grid search on the development set. For the English monolingual setting (en-en), the best performance was achieved by assigning a weight of 0.9 to mE5-large-instruct and 0.1 to mE5-large. In contrast, the Spanish monolingual setting (es-es) benefited from a more balanced combination, with weights of 0.6 and 0.4, respectively. These results suggest that the two models capture complementary semantic signals, and that the relative contribution of each model varies across languages. Overall, the ensemble achieved MAP scores of 0.9000 on English and 0.9031 on Spanish, outperforming all individual models. For the Cross-Lingual (en-es) setting, this weight optimization procedure could not be replicated locally, as development-set performance for this setting could not be obtained. Only mE5-large-instruct was used without ensemble for this setting.

To assess the statistical significance of the ensemble's improvement over individual models, a Wilcoxon signed-rank test was conducted on per-query Average Precision scores on the development set. For the English setting, mE5-large-instruct significantly outperformed the baseline ($p = 0.0020$), while the ensemble's improvement over mE5-large-instruct alone was not statistically significant ($p = 0.3262$), consistent with the marginal MAP gain observed. A similar pattern was observed for Spanish: mE5-large-instruct significantly outperformed the baseline ($p = 0.0010$), while the ensemble's improvement over mE5-large-instruct was not statistically significant ($p = 0.1250$). The ensemble nonetheless significantly outperformed the baseline in both languages ($p = 0.0020$ and $p = 0.0010$).

It should be noted that the Task A development set contains only 10 queries per language, which limits the statistical power of these tests. The non-significant ensemble-vs-instruct comparisons should therefore be interpreted as inconclusive rather than as evidence of no true difference.

Table 5 further compares the proposed method with the official TalentCLEF baseline across all three submitted settings on the test set. On the test set, the proposed system achieved MAP scores of 0.6327 for English (en-en) and 0.6202 for Spanish (es-es), substantially outperforming the official baseline, which obtained 0.4001 and 0.3655, respectively. This corresponds to relative improvements of 58.1% on English and 69.7% on Spanish. For the Cross-Lingual (en-es) setting, the system achieved a MAP of 0.5995, compared to the baseline score of 0.3562, representing a relative improvement of 68.3%. These results demonstrate that the proposed multilingual dense retrieval framework generalizes effectively to

unseen data and remains robust across all three evaluation settings. The Bias-Controlled test results are discussed in Section 5.4

Table 5
Comparison between the official baseline and the proposed ensemble on Task A.

Method	Setting	MAP (Development)	MAP (Test)
TalentCLEF Baseline	en-en	0.7060	0.4001
	es-es	0.6470	0.3655
	en-es	—	0.3562
Proposed	en-en	0.9000	0.6327
	es-es	0.9031	0.6202
	en-es	—	0.5995

5.2. Task B Results

Table 6 presents the evaluation results on the Task B development set. Following the TalentCLEF 2026 guidelines, nDCG is used as the primary evaluation metric under both Binary Relevance and Graded Relevance settings. MAP is additionally reported to provide a more comprehensive assessment of retrieval effectiveness.

Table 6
Task B validation set results.

Method	nDCG (Binary)	nDCG (Graded)	MAP
all-MiniLM-L6-v2 (baseline)	0.6537	0.6107	0.1360
mE5-large-instruct (zero-shot)	0.7044	0.6606	0.1956
Hybrid semantic lookup	0.7081	0.6683	0.2005
mE5-large-instruct (fine-tuned)	0.7094	0.6657	0.2117
Ensemble (proposed)	0.7341	0.6905	0.2437

Overall, all variants based on mE5-large-instruct outperformed the official baseline (all-MiniLM-L6-v2) across all evaluation metrics. The zero-shot version of mE5-large-instruct achieved an nDCG (binary) of 0.7044, an nDCG (graded) of 0.6606, and a MAP of 0.1956, compared to 0.6537, 0.6107, and 0.1360 obtained by the baseline, respectively. These results indicate that multilingual retrieval-oriented embeddings are effective for modeling semantic relationships between job titles and skills without task-specific fine-tuning.

The Hybrid Semantic Lookup approach, which combines dense retrieval with the official occupation-to-skill mapping, yielded a modest improvement over the zero-shot model. The method achieved an nDCG (binary) of 0.7081, an nDCG (graded) of 0.6683, and a MAP of 0.2005. This suggests that structured knowledge from occupation-skill relationships can provide additional relevance signals, although the overall contribution remains limited compared to semantic embeddings.

Fine-tuning mE5-large-instruct on 5,000 essential job-skill pairs further improved performance, particularly in terms of MAP. The fine-tuned model achieved a MAP of 0.2117, compared to 0.1956 for the zero-shot model. However, improvements in nDCG were relatively small, increasing from 0.7044 to 0.7094 for binary relevance and from 0.6606 to 0.6657 for graded relevance. This observation suggests that fine-tuning helps the model retrieve a larger number of relevant skills overall, but does not substantially improve the ranking of the most relevant skills at the top positions.

The fine-tuning configuration described in Section 4.2.3 was adopted following initial attempts at higher learning rates (1×10^{-5} and 2×10^{-5}) on the full essential-pair set, which caused catastrophic forgetting and resulted in nDCG dropping to approximately 0.49 and MAP to approximately 0.02, well below the zero-shot baseline of 0.7044 and 0.1956, respectively. Reducing the learning rate to 2×10^{-6}

and limiting the training sample to 5,000 pairs successfully mitigated this issue, recovering nDCG to 0.7094 and MAP to 0.2117. Nevertheless, the extent to which performance is further limited by training data quantity or additional hyperparameter tuning remains an open question for future work.

The best overall performance on the development set was obtained using the proposed ensemble strategy, which combines the zero-shot and fine-tuned versions of mE5-large-instruct. In both settings, job titles were encoded using the task-specific instruction prefix *"Given a job title, retrieve the most relevant skills for this job"*, while the skill corpus was expanded using all available ESCO aliases to maximize semantic coverage and improve the retrieval of skill variants expressed with different terminology.

The fine-tuned model was obtained by further training mE5-large-instruct on 5,000 randomly sampled essential job-skill pairs from the official occupation-to-skill training data using `MultipleNegativesRankingLoss`. While the zero-shot model benefits from strong multilingual semantic representations learned during pretraining, the fine-tuned model incorporates additional domain-specific knowledge regarding occupation-skill relationships.

To combine the strengths of both models, retrieval scores were first normalized using min-max normalization on a per-query basis and then merged using weighted score fusion. The optimal weights were determined through grid search on the development set, resulting in a weight of 0.55 for the zero-shot model and 0.45 for the fine-tuned model. The slightly higher weight assigned to the zero-shot model suggests that its pretrained semantic knowledge remains the primary source of retrieval effectiveness, while the fine-tuned model provides complementary domain-specific signals that further improve ranking quality.

The resulting ensemble achieved an nDCG (binary) of 0.7341, an nDCG (graded) of 0.6905, and a MAP of 0.2437, outperforming all other methods. These findings indicate that the zero-shot and fine-tuned models capture different aspects of relevance and that combining them leads to more effective job-skill matching than using either model individually.

Table 7 further compares the proposed method with the official TalentCLEF baseline on the test set.

Table 7
Task B results on the validation and official test sets.

Method	Split	nDCG (Binary)	nDCG (Graded)	MAP
TalentCLEF Baseline	Dev	0.6537	0.6107	0.1360
	Test	0.6634	0.6204	0.1471
Proposed	Dev	0.7341	0.6905	0.2437
	Test	0.7278	0.6821	0.2366

On the test set, the proposed ensemble achieved an nDCG (binary) of 0.7278, an nDCG (graded) of 0.6821, and a MAP of 0.2366, substantially outperforming the official baseline, which achieved 0.6634, 0.6204, and 0.1471, respectively. This corresponds to relative improvements of 9.7% for nDCG (binary), 9.9% for nDCG (graded), and 60.8% for MAP. Overall, the results demonstrate that the combination of multilingual dense retrieval, domain-specific fine-tuning, and score-based ensemble strategies is effective for job-skill matching. The consistent improvements observed on both the development and test sets suggest that the proposed approach generalizes well to unseen occupations and skills.

5.3. Error Analysis

Error analysis was conducted on the proposed models that achieved the highest performance across both Task A and Task B. Although aggregate evaluation metrics offer a general measure of retrieval effectiveness, they do not fully capture how the model behaves on individual queries. Therefore, a detailed examination of representative best-performing and worst-performing queries helps clarify the strengths and limitations of the proposed approach. For Task A, the analysis relies on query-level Average Precision (AP) scores on the development set. For Task B, the evaluation also factors in binary and graded nDCG scores, the number of relevant skills, and Precision@5 (P@5), which measures the

proportion of relevant skills among the top five retrieved results. This analysis yields deeper insights into the specific query types that the model processes successfully and those that continue to pose challenges.

5.3.1. Task A Error Analysis

Table 8 displays a selection of queries from the Task A development set that yielded the highest and lowest results. The proposed system maintained strong outcomes for every analyzed query because the minimum Average Precision (AP) remained above 0.73. This outcome indicates that the retrieval system successfully located appropriate candidate resumes even though job descriptions and resumes contained intricate and varied details.

Table 8
Representative queries used for error analysis on the Task A development set.

Job Description	Language	AP
Senior Mechanical/HVAC Engineer	English	0.7497
Ingeniero/a Mecánico/a Sénior	Spanish	0.7353
Healthcare Data Analyst	English	0.9926
Analista de Datos de Salud	Spanish	0.9789
IELTS Teacher	English	1.0000
Profesor/a de IELTS	Spanish	1.0000

Within the weaker results, *Senior Mechanical/HVAC Engineer* recorded an AP score of 0.7497, while its Spanish equivalent, *Ingeniero/a Mecánico/a Sénior*, scored 0.7353. These figures are still high but fall below the scores of the remaining queries. Engineering roles often demand a broad mix of specialized certifications, technical capabilities, and field experience. Consequently, numerous resumes exhibit comparable credentials. This overlap hinders the ability of the model to separate the optimal candidate from profiles that are only partially qualified.

Conversely, *Healthcare Data Analyst* reached an AP score of 0.9926, and *Analista de Datos de Salud* reached 0.9789. Furthermore, *IELTS Teacher* and *Profesor/a de IELTS* both secured a perfect AP score of 1.0000. These outcomes indicate that the target resumes feature unique credentials and skill sets that align perfectly with the job criteria. Therefore, the system placed the correct resumes at the top of the ranking with little to no uncertainty.

The steady search performance across different languages represents a noteworthy finding. For instance, the difference between *Healthcare Data Analyst* and *Analista de Datos de Salud* is a mere 0.0137 AP points. Meanwhile, *IELTS Teacher* and *Profesor/a de IELTS* generated identical scores. This trend demonstrates that the multilingual retrieval system successfully maps parallel semantic representations for English and Spanish. As a result, the model delivers equivalent retrieval accuracy for job descriptions that share identical meanings.

The representative performance of the proposed framework indicates that retrieval effectiveness often declines when processing roles that demand diverse and varied technical skills, as several candidates frequently present comparable qualifications. Regardless of this trend, the high Average Precision (AP) scores sustained across both English and Spanish queries confirm that the proposed multilingual dense retrieval framework remains highly robust for contextualized job-person matching.

5.3.2. Task B Error Analysis

Table 9 displays a selection of queries from the Task B development set that produced the highest and lowest outcomes. Unlike Task A, the performance variation among these queries is much wider, with AP scores spanning from 0.0069 to 0.5551. This divergence proves that the difficulty of matching jobs with skills shifts dramatically across different professions, largely depending on how clearly a job title links to a specific set of competencies.

Table 9
Representative queries used for error analysis on the Task B development set.

Job Title	AP	nDCG (Binary)	nDCG (Graded)	#Core	#Contextual	P@5
GL Specialist	0.0069	0.3418	0.3059	16	32	0.0
SDET II	0.0296	0.4997	0.4564	22	112	0.2
Operations Processor III	0.0548	0.5768	0.5391	46	116	0.4
Lead Electronics Technician	0.4418	0.8615	0.8268	86	233	0.6
Licensed Nurse	0.5551	0.8892	0.8476	101	212	0.8

The weakest performance came from the *GL Specialist* query, which recorded an AP of 0.0069, a binary nDCG of 0.3418, and a graded nDCG of 0.3059. Additionally, its P@5 stood at 0.0, meaning the system failed to place a single relevant skill within the top five results. Although the ground truth contained 16 core skills and 32 contextual skills, the framework failed to extract them successfully. This low accuracy likely stems from the ambiguity of the abbreviation "GL", which can signify multiple distinct concepts, thereby preventing the system from forming a precise semantic bridge between the job title and the correct skills.

In a similar manner, *SDET II* and *Operations Processor III* also yielded weak AP scores of 0.0296 and 0.0548, respectively. Both designations incorporate acronyms, seniority levels, or internal corporate jargon that fails to offer enough semantic context for the retrieval system. Their P@5 values of 0.2 and 0.4 confirm that only a fraction of the highest-ranked skills matched the actual requirements.

Conversely, *Licensed Nurse* led the analyzed queries with an AP of 0.5551, a binary nDCG of 0.8892, a graded nDCG of 0.8476, and a P@5 of 0.8. Similarly, *Lead Electronics Technician* secured an AP of 0.4418 alongside a P@5 of 0.6. These specific roles correlate strongly with highly standardized skill sets, allowing the framework to identify and organize relevant competencies with greater accuracy. Furthermore, both positions connect to an extensive array of valid skills, including 101 core and 212 contextual skills for *Licensed Nurse*, and 86 core and 233 contextual skills for *Lead Electronics Technician*.

In summary, the findings show that retrieval accuracy improves for professions that maintain explicit, standardized connections to specific competencies. On the other hand, effectiveness drops when job titles are vague, shortened, or unique to a single company, because these traits hinder the framework from building precise semantic representations. This pattern remains visible in the P@5 metric, where top-performing queries regularly position a greater share of appropriate skills at the top of the ranking compared to low-performing queries.

5.4. Limitations

This study has one primary limitation that should be acknowledged. Although the system was submitted to the Bias-Controlled evaluation setting in Task A, which measures the ability of a retrieval system to minimize performance disparities across different demographic groups, no dedicated bias analysis or mitigation strategy was applied. The submitted system achieved an AVG_RBO of 0.8040 compared to the official baseline score of 0.7822. A thorough analysis of retrieval performance disparities across demographic groups, as well as the development of dedicated mitigation strategies, remains an important direction for future work.

6. Conclusion and Future Work

This paper presented a multilingual bi-encoder retrieval system for TalentCLEF 2026, evaluated on two tasks: Contextualized Job-Person Matching (Task A) and Job-Skill Matching with Skill Type Classification (Task B). The proposed system employs `intfloat/multilingual-e5-large-instruct` as the primary retrieval model, combined with a score-based ensemble strategy to further improve performance.

Experimental results demonstrate that the proposed approach consistently outperforms the official TalentCLEF 2026 baselines on both tasks. For Task A, the ensemble achieved MAP scores of 0.6327, 0.6202, and 0.5995 on the English monolingual, Spanish monolingual, and Cross-Lingual (en-es) test sets, corresponding to relative improvements of 58.1%, 69.7%, and 68.3% over the baseline, respectively. For Task B, the ensemble obtained an nDCG (binary) of 0.7278, an nDCG (graded) of 0.6821, and a MAP of 0.2366 on the test set, representing improvements of 9.7%, 9.9%, and 60.8% over the baseline, respectively. These findings highlight the effectiveness of combining multilingual instruction-tuned embeddings with ensemble strategies for HCM retrieval tasks.

Future work for Task A could explore fairness-aware retrieval approaches to better address the Bias-Controlled evaluation setting, as well as ensemble strategies with architecturally distinct models. Future work for Task B could include query expansion or entity disambiguation techniques for ambiguous and abbreviated job titles, scaling fine-tuning to a larger portion of the available training data combined with a systematic hyperparameter search, and applying a similarly diverse ensemble approach.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude, ChatGPT, and Gemini in order to: drafting content, text translation, grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] 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.

[2] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, R. Zbib, Overview of the talentclef 2025: Skill and job title intelligence for human capital management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2025, pp. 464–485.

[3] H. Fabregat, L. García-Sardiña, P. Estrella, L. Gasco, C. P. Carrino, D. Deniz Cerpa, A. Rodrigo, R. Zbib, Overview of talentclef 2026: Task a — contextualized job-person matching, in: CLEF (Working Notes), 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 (Working Notes), 2026.

[5] M.-H. Ajjam, H. S. Al-Raweshidy, Ai-driven semantic similarity-based job matching framework for recruitment systems, Information Sciences 724 (2026) 122728. URL: https://www.sciencedirect.com/science/article/pii/S0020025525008643.

[6] G. Xie, C. Tan, Design of intelligent employment recommendation system for agricultural industry based on twin network and bert, in: Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education, ICAIE '24, Association for Computing Machinery, New York, NY, USA, 2025, p. 862–867. URL: https://doi.org/10.1145/3722237.3722386.

[7] J. Rosenberger, L. Wolfrum, S. Weinzierl, M. Kraus, P. Zschech, Careerbert: Matching resumes to esco jobs in a shared embedding space for generic job recommendations, Expert Systems with Applications 275 (2025) 127043. URL: https://www.sciencedirect.com/science/article/pii/S0957417425006657.

[8] M. B. Myneni, F. S. Quadhari, Hybrid transformer-based resume parsing and job matching using textrank, sbert, and deberta, International Journal of Electronics and Communication Engineering 12 (2025) 63–71. doi:10.14445/23488549/IJECE-V12I9P105.

[9] R. Alonso, D. Dessí, A. Meloni, D. Reforgiato Recupero, A novel approach for job matching and

skill recommendation using transformers and the o*net database, Big Data Research 39 (2025) 100509. URL: https://www.sciencedirect.com/science/article/pii/S2214579625000048.

[10] M. Herold, M. Roedenbeck, Matching competency frameworks with job advertisements: a data-driven analysis of its practical application in the healthcare sector, Evidence-based HRM: a Global Forum for Empirical Scholarship 13 (2024) 339–356. URL: https://doi.org/10.1108/EBHRM-07-2023-0181.

[11] M. H. Nizami, A. Uddin, M. T. Salani, A. Saeed, F. Alvi, A. Samad, Enhancing human capital management: AI techniques for candidate matching and skill extraction, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_371.pdf.

[12] A. Tejera Villar, I. Segura Bedmar, HULAT-UC3M at TalentCLEF: Artificial intelligence and natural language processing applied to HR management, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_374.pdf.

[13] M. Zhang, R. van der Goot, NLPnorth @ TalentCLEF 2025: Comparing discriminative, contrastive, and prompt-based methods for job title and skill matching, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_377.pdf.

[14] R. Barakat, O. Mokhtar, M. Torki, N. Elmakky, AlexU-NLP at TalentCLEF 2025: Curriculum-driven hybrid retrieval for multilingual job title matching, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_364.pdf.

[15] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.pdf.

[16] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 296–310. URL: https://aclanthology.org/2021.naacl-main.28.pdf.

[17] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).

[18] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W.-t. Yih, N. A. Smith, L. Zettlemoyer, T. Yu, One embedder, any task: Instruction-finetuned text embeddings, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 1102–1121. URL: https://aclanthology.org/2023.findings-acl.71/.

[19] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive text embedding benchmark, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 2014–2037. URL: https://aclanthology.org/2023.eacl-main.148/.

[20] L. Gasco, H. Fabregat Marcos, P. Estrella, L. García-Sardiña, C. P. Carrino, D. Deniz, R. Zbib, J.-J. Decorte, M. De Lange, W. Veys, A. Rodrigo, Talentclef 2026 corpus: Skill and job title intelligence for human capital management, 2026. URL: https://doi.org/10.5281/zenodo.19652670. doi:10.5281/zenodo.19652670.