CYUT at TalentCLEF 2026: Zero-Shot Multi-View Retrieval for Job-Person Matching

Notebook for the TalentCLEF Lab at CLEF 2026

Johan Poerwanto[1], Shih-Hung Wu[1,*]

[1]Chaoyang University of Technology, Taichung, Taiwan (R.O.C.)

Abstract

Job-person matching can be approached through various retrieval methods, but no single method dominates across all cases. Dense retrieval captures semantic relationships but misses exact lexical matches, while sparse retrieval handles specific terminology well but struggles with paraphrased content. These challenges are compounded in multilingual settings. In this paper, we present Multi-View Reciprocal Rank Fusion (MV-RRF), a multi-view retrieval system that fuses five ranked lists through weighted RRF: dense retrieval using BGE-M3, forward Hypothetical Document Embeddings (HyDE) that generates hypothetical CV profiles from job descriptions, a concatenated dense variant over the job plus HyDE text, BM25 sparse retrieval, and reverse HyDE that generates hypothetical job descriptions from candidate CVs. We further study taxonomy-grounded enrichment of forward HyDE using occupational taxonomies (ESCO, O*NET). Fused rankings are refined with listwise reranking (RankZephyr). Our experiments show that weighted RRF effectively combines dense and sparse views, that injecting domain-level occupational taxonomy into forward HyDE reliably improves performance over plain HyDE, but that fusion with a strong BM25 route largely attenuates this gain because lexical matching dominates and the taxonomy signal is concentrated in the HyDE route.

Keywords

Job-Person Matching, Reciprocal Rank Fusion, Hypothetical Document Embeddings, Listwise Reranking, Multilingual Information Retrieval, Occupational Taxonomies, BGE-M3, BM25

1. Introduction

The task of matching job descriptions to candidate profiles is a fundamental challenge in human capital management. TalentCLEF [1] is a recurring evaluation forum for skill and job title intelligence in this domain. Its 2026 edition [2] formalizes job-person matching as a multilingual ranking problem covering English and Spanish, with evaluation based on Mean Average Precision (MAP). Given a job description as query, a system must rank a pool of candidate CVs by relevance. This retrieval problem is complicated by the asymmetries between how jobs describe desired qualifications and how candidates describe past experience.

Recent work has shown that learning-to-retrieve approaches can be highly effective for job matching. Shen et al. [3] demonstrated this at LinkedIn's scale by leveraging confirmed hires and millions of user interactions to train retrieval models. However, this approach requires access to large volumes of real behavioral data such as click-through rates, application outcomes, and hiring decisions, which most organizations do not have. Moreover, real recruitment data carries significant privacy constraints, making it difficult to share or reproduce across research settings. This raises a practical question: how far can we push job-person matching without any task-specific training data?

Two dominant retrieval methods offer complementary strengths. Sparse methods such as BM25 rely on exact lexical matching and are effective when job descriptions and CVs share specific terminology, but fail when faced with semantically equivalent terms. Dense retrieval encodes texts into continuous vector representations and matches by semantic similarity, handling paraphrasing well but tending to blur distinctions between related-but-different roles. Tonellotto and Macdonald [4] demonstrated

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Corresponding author.

s11330655@cyut.edu.tw (J. Poerwanto); shwu@cyut.edu.tw (S. Wu)

that sparse and dense retrievers retrieve complementary sets of relevant documents, motivating hybrid approaches.

To bridge the asymmetries between job descriptions and CVs without training data, we adopt Hypothetical Document Embeddings (HyDE) [5], generating synthetic candidate profiles from job descriptions for retrieval. We then add reverse HyDE, which generates hypothetical job descriptions from CVs, inspired by the reverse hypothetical prompt embedding approach of Mackie and Dalton [6], and study taxonomy-grounded enrichment of forward generation by injecting structured occupational context from ESCO and O*NET. These multiple views are combined through weighted Reciprocal Rank Fusion [7] and refined with listwise reranking.

In this paper, we present MV-RRF (Multi-View Reciprocal Rank Fusion), a training-free job-person matching system that fuses all five routes through weighted RRF and applies optional listwise reranking. Supervised matchers and direct LLM scorers (Section 2) require labels or pairwise scoring at scale; on the TalentCLEF 2026 Task A development split we instead ask where zero-shot fusion gains come from. We compare taxonomy-grounded forward HyDE against a plain job-description baseline to test whether occupational context improves generation; ablate individual routes and fusion subsets to see which views survive alongside a dominant BM25 list; tune per-route RRF weights to separate fusion design from weight choice; and evaluate SlideGar listwise reranking on English and Spanish before submitting our final configuration to the official test benchmark across `en-en`, `es-es`, and cross-lingual `en-es`.

2. Related Work

2.1. Person-Job Fit

The task of matching candidates to job positions has been modeled primarily as a supervised text matching problem. Zhu et al. [8] proposed Person-Job Fit Neural Network (PJFNN), using hierarchical Convolutional Neural Network (CNN) to jointly represent resumes and jobs, while Yang et al. [9] introduced a dual-perspective graph neural network that models two-way selection preference between candidates and employers.

More recently, Yu et al. [10] showed that dense retrieval with contrastive learning achieves strong results without complex architectures. Their follow-up, ConFit V2 [11], introduced Hypothetical Resume Embedding (HYRE). This approach uses an LLM to generate a hypothetical resume from a job post to augment the job representation, alongside a runner-up hard-negative mining strategy. HYRE is conceptually related to our use of forward HyDE, though ConFit V2 requires labeled accept/reject pairs for contrastive training, while our approach operates in a fully zero-shot setting without any relevance labels.

Kostis et al. [12] explored an integrated retrieval system using ESCO-based representations to semantically match CVs, job descriptions, and curricula. Our taxonomy-grounded enrichment extends this direction by using ESCO and O*NET concepts to enrich LLM-based query generation rather than as direct matching features.

2.2. LLM-Based Matching

Large language models have been explored for direct resume-job scoring. Vaishampayan et al. [13] compared zero-shot GPT-4 ratings against human judgments for 736 resumes and found that LLM scores correlate only weakly with human assessments. Additionally, direct LLM scoring requires processing each job-CV pair individually, making it expensive at scale. Rather than using LLMs as direct matchers, our approach uses them for two bounded generation tasks: forward and reverse HyDE, after which standard embedding-based retrieval handles the ranking.

2.3. Zero-Shot Retrieval

Thakur et al. [14] established that zero-shot retrieval performance varies significantly across domains, highlighting the need for domain-specific adaptations even in training-free settings. HyDE addresses this by using LLM generation to bridge the query-document gap without any relevance labels. Our work extends HyDE with reverse generation and taxonomy-grounded enrichment, and combines it with sparse retrieval through rank fusion. This targeting should address the job-CV asymmetry in recruitment.

3. Dataset

The Task A portion of the TalentCLEF 2026 corpus [15] contains job postings and candidate CVs released on Zenodo under a Creative Commons Attribution 4.0 license. The dataset includes separate English and Spanish materials, spanning multiple job domains and professional sectors. Systems treat each job posting as a query and rank CV documents from the pool provided for that language.

- **Development data.** For each language, the development split pairs job postings with a CV pool for retrieval and local evaluation. English and Spanish share the same split size: 10 queries and 472 CVs per language.
- **Test data.** For each language, the test split supplies job postings and a CV pool for building ranked lists to submit to the organizers. English and Spanish share the same split size: 40 queries and 476 CVs per language; official scoring is applied to these test queries [15].

Note. Query-document relevance judgments are available for `en-en` and `es-es` on the development split, enabling local evaluation and comparison to the official Codabench development baseline (Table 4). For *cross-lingual* matching (`en-es`), no relevance judgments are provided for the development split.

4. Methodology

In this section, we first formally describe the task formulation for Task A. We then describe the five retrieval routes used in our system in Figure 1. Finally, we describe Reciprocal Rank Fusion (consensus and weighted variants) and listwise reranking used to combine the retrieval results. Following the demand for cross-lingual evaluation, we also describe our multilingual matching setup.

4.1. Task Formulation

Task A is Contextualized Job-Person Matching. Given a job description q as query, a system must rank a corpus of N candidate CVs $\{d_1, d_2, \ldots, d_N\}$ by relevance. The task covers three language modes: monolingual English (en-en), monolingual Spanish (es-es), and cross-lingual English-to-Spanish (en-es). The primary evaluation metric is Mean Average Precision (MAP); dataset statistics are given in Section 3.

4.2. Retrieval Routes

Our system produces five independent ranked lists per query (routes A through E), each capturing a different aspect of the job-person fit.

4.2.1. Route A: Dense Retrieval

Route A embeds both the job description q and all candidate CVs using BGE-M3 [16], a multilingual bi-encoder supporting over 100 languages. Documents are ranked by cosine similarity between L2-normalized embeddings. We use a maximum sequence length of 512 tokens.

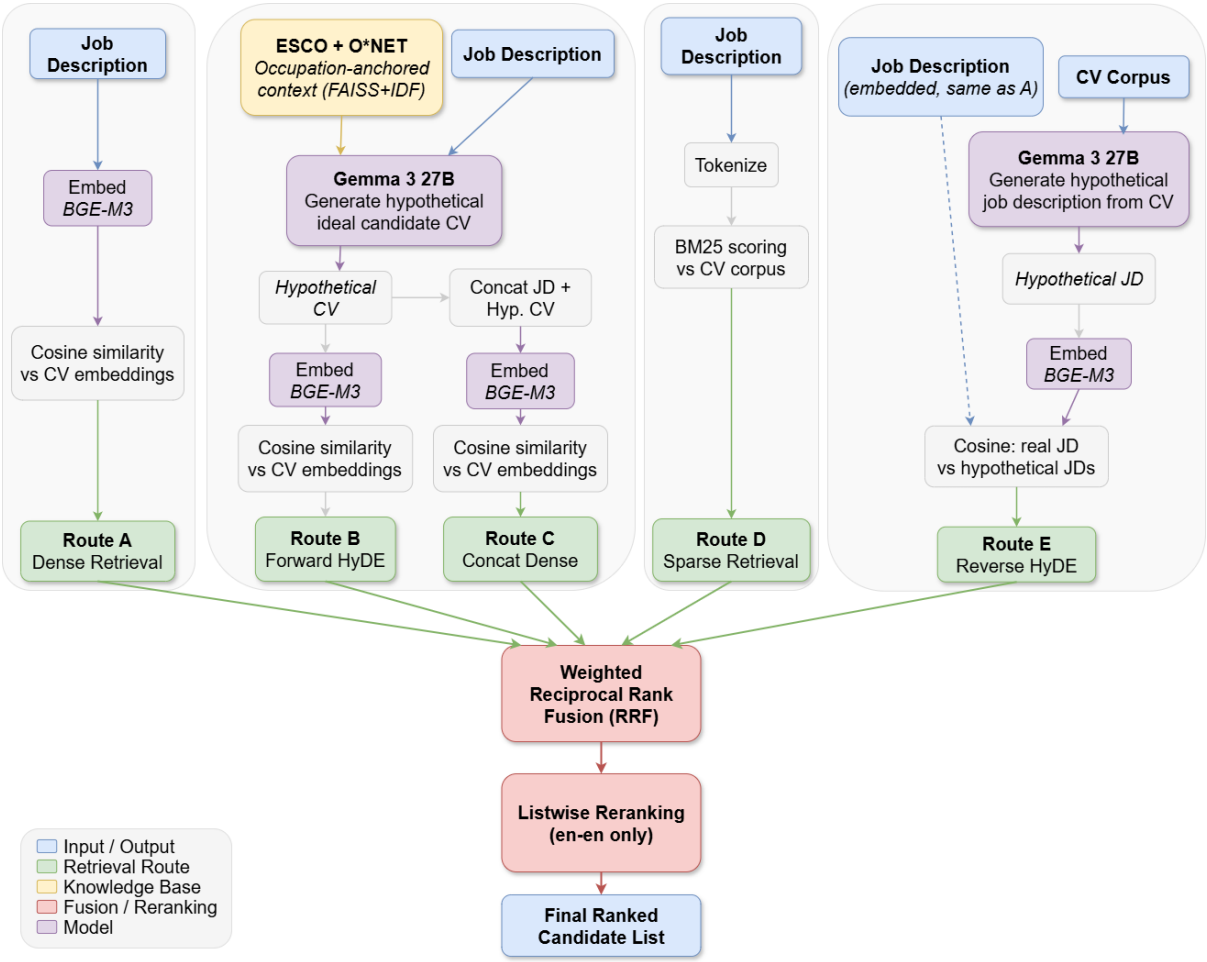

Figure 1: Overview of the MV-RRF system: five retrieval routes (A–E) produce ranked candidate lists that are fused with weighted RRF; listwise reranking refines the fused list.

4.2.2. Route B: Forward HyDE with Taxonomy-Grounded Enrichment (TaxHyDE)

Forward HyDE generates a hypothetical ideal candidate CV from the job description using **Gemma3:27B** [17]. The generated profile is then embedded with BGE-M3 and ranked against the CV corpus by cosine similarity, following the approach introduced by Gao et al. [5].

To improve the specificity of the generated profile, we enrich the generation prompt with structured occupational context from two complementary taxonomies: ESCO v1.2.0[1] and O*NET v30.2[2]. We term this variant **TaxHyDE** (Taxonomy-grounded HyDE). Without taxonomy snippets, the same generator uses the job description only (*plain HyDE*). We evaluate that baseline in Section 5.

The *source* of taxonomy context is itself an experimental variable. Beyond plain HyDE and the occupation-anchored variant described here, we compare drawing context from ESCO only, O*NET only, both combined, and a direct IDF-retrieval variant; these B-context modes are ablated in Section 5.

The enrichment follows an occupation-anchored two-hop retrieval process:

1. **Taxonomy index construction (offline).** We parse occupations, skills, tasks, tools, and knowledge elements from both ESCO and O*NET into a unified record set (∼44K entries). Each record is embedded with BGE-M3 and indexed in Facebook AI Similarity Search (FAISS) [18] `IndexFlatIP` for nearest-neighbor retrieval.
2. **First hop: occupation retrieval.** The full job description is embedded and queried against the taxonomy index, filtered to occupation-type records only. The top-k most similar occupations

[1] https://esco.ec.europa.eu/
[2] https://www.onetcenter.org/

are retrieved.

3. **Second hop: linked concept lookup.** For each retrieved occupation, all skills, tasks, and tools linked to it in the taxonomy metadata are collected into a candidate concept pool.
4. **IDF ranking.** Each concept in the pool is weighted by its Inverse Document Frequency across the CV corpus, computed using semantic similarity rather than substring matching to handle multilingual content. Concepts are split into two tiers: critical (high IDF, rare and specialized) and common (low IDF, broadly appearing) and formatted into the generation prompt.

The motivation for combining ESCO and O*NET is that they provide complementary granularity: ESCO offers broad multilingual skill labels that match CV vocabulary, while O*NET provides task-level descriptions with specific technical terminology (e.g., "Perform electron microscopy to analyze specimens").

4.2.3. Route C: Concatenated Dense Retrieval

Route C concatenates the job description with the forward HyDE profile from Route B into a single text, embeds it with BGE-M3, and ranks by cosine similarity against the CV corpus. Route B ranks by the generated profile alone. Route C keeps the original query text in the same embedding. The same B-context supplies the HyDE profile for both routes.

4.2.4. Route D: Sparse Retrieval

Route D uses BM25 [19] to score candidate CVs against the raw job description text. This captures exact lexical matches that dense retrieval may overlook, particularly for domain-specific abbreviations and technical terms.

4.2.5. Route E: Reverse HyDE

Reverse HyDE generates a hypothetical job description from each candidate CV using **Gemma3:27B**. These hypothetical JDs are pre-embedded with BGE-M3 offline. At query time, the real job description embedding (the same vector as Route A) is compared against the pre-embedded hypothetical JDs by cosine similarity.

This route bridges the JD-CV asymmetry from the opposite direction as Route B. While forward HyDE asks "what would the ideal candidate look like?", reverse HyDE asks "what job would this candidate apply to?" The two directions capture complementary matching signals, inspired by the hypothetical prompt embedding approach of Mackie and Dalton [6].

4.3. Reciprocal Rank Fusion

Active routes produce independent ranked lists that are fused before reranking. We use two variants: *consensus RRF* for development ablations (condition C_1 in Section 5) and *weighted RRF* for submission tuning (condition C_2). Both build on standard reciprocal rank fusion [7] with smoothing constant k=60.

Consensus RRF. For each candidate document d, let $R_d \subseteq R$ be the active routes whose ranked lists contain d, and let $n = |R_d|$. The fused score is

$$\text{score}_{\text{cons}}(d) = \left(1 + \frac{n}{10}\right) \sum_{r \in R_d} \frac{1}{k + \text{rank}_r(d)} \tag{1}$$

The factor $(1 + n/10)$ is a *consensus bonus*: documents retrieved by more routes receive a multiplicative boost on top of the usual RRF sum. This follows the Exp4Fuse-style adaptive weighting of Liu and Zhang [20]. We apply it over up to five route lists rather than two query expansions. All routes contribute equally inside the sum (implicit unit weight).

Weighted RRF. For submission and weighted tuning we omit the consensus multiplier and assign a tunable weight w_r per route:

$$\text{score}_{\text{w}}(d) = \sum_{r \in R_d} \frac{w_r}{k + \text{rank}_r(d)} \tag{2}$$

where R is the set of active routes, w_r defaults to 1.0 in grid search, and R_d is defined as above. Equal weights are therefore not identical to consensus RRF because of the missing $(1 + n/10)$ term. Route sets and weights are selected on the development set (Section 5).

4.4. Listwise Reranking

The top candidates from RRF fusion are refined using listwise reranking with RankZephyr 7B [21], a Zephyr-based model fine-tuned for listwise passage ranking. We adopt a sliding window approach inspired by SlideGar [22]. A window of w documents is presented to the reranker, which produces a permutation. The top b documents carry forward to the next window, and the remaining documents are placed in the final output. This process repeats until a budget of c documents have been reranked. The parameters w, b, and c are tuned on the development set.

SlideGar optionally supports corpus graph expansion, where nearest-neighbor CVs (precomputed from BGE-M3 embeddings) are injected into subsequent sliding windows to discover relevant candidates that may have been missed by first-stage retrieval. We evaluate this option in our experiments.

Since RankZephyr is trained on English data (MS MARCO), an additional translation step is required when reranking non-English documents. For Spanish, the top-c CVs and the query are translated to English prior to reranking, and the resulting permutation is mapped back to the original Spanish document IDs.

4.5. Cross-Lingual Matching

For the English-to-Spanish evaluation (en-es), we adopt an English-anchor strategy: all Spanish CVs are translated to English prior to indexing. The translated CVs are then processed through the same retrieval pipeline as the English monolingual setting, allowing all routes, including BM25, which requires lexical overlap and cannot function cross-lingually on untranslated text, to operate in a shared language space.

The choice of translation model affects downstream quality, as CVs are semi-structured documents where section headers, company names, and date formatting carry matching signal. We compare sentence-level and document-level translation approaches in our experiments.

5. Experiments

5.1. Experimental Setup

We evaluate on the TalentCLEF 2026 Task A dataset [15], described in Section 3. No relevance labels are provided for the test set; evaluation is performed through submission to the Codabench platform.

All dense retrieval and embedding operations use BGE-M3 [16] with a maximum sequence length of 512 tokens. HyDE generation (forward and reverse) uses **Gemma3:27B** [17] served via Ollama. Listwise reranking uses RankZephyr 7B [21]. The taxonomy index is built from ESCO v1.2.0 and O*NET v30.2, comprising approximately 44K records embedded with BGE-M3 in the FAISS index described above. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Table 1 collects the notation, route labels, and condition definitions used throughout the paper. Each results table is produced under one of two registered conditions (Table 2) and repeats its setup in a header row. A condition fixes three things at once—the B-context, the fusion method, and the route set—so consensus versus weighted RRF is only the *fusion* axis of a condition, not the condition itself. Routes A, D, and E are fixed retrieval mechanisms; B and C always use the B-context active in that run.

Table 1
Nomenclature used throughout the paper: retrieval routes, B-context modes, notation convention, fusion variants, and the two registered conditions.

Symbol / term	Category	Meaning
A	Route	Dense retrieval; BGE-M3 over the raw job description.
B	Route	Forward HyDE; generated ideal-CV profile from the job description.
C	Route	Concatenated dense; job description + forward HyDE profile embedded together.
D	Route	Sparse retrieval; BM25 over the raw job description.
E	Route	Reverse HyDE; generated job description from each candidate CV.
plain HyDE	B-context	No taxonomy; job-description-only generation prompt.
ESCO / O*NET / Both	B-context	Forward-HyDE prompt enriched from that taxonomy source.
`faiss-idf`	B-context	Direct IDF retrieval over the full 44K index; bypasses occupation matching.
occupation-anchored (TaxHyDE)	B-context	Two-hop occupation→concept retrieval, IDF-ranked (submitted system).
$B_{\text{esco}}, C_{\text{esco}}$	Notation	Route B/C on raw single-source context; subscript = *source* label.
$B^{\text{anch}}, C^{\text{anch}}$	Notation	Route B/C on occupation-anchored TaxHyDE; superscript "anch" = *method* label.
consensus RRF	Fusion	Unit-weight RRF with the $(1+n/10)$ consensus bonus.
weighted RRF	Fusion	Per-route tunable weights; no consensus bonus.
C_1	Condition	Raw B-context + consensus RRF + full ablation (see Table 2).
C_2	Condition	B^{anch} + weighted RRF + fixed submission route sets (see Table 2).

Table 2
The two experimental conditions, each fixing a B-context, a fusion method, and a route-set policy (development set). B_{esco}: raw ESCO B-context (Tables 3–4). B^{anch}: occupation-anchored two-hop TaxHyDE (Section 4).

ID	B-context	Fusion	Route-set policy	Tables
C_1	raw B-context (ESCO/O*NET/Both)	consensus RRF	full ablation	3, 4
C_2	B^{anch}	weighted RRF	EN: A, B, D, E; ES: C, D, E	5, 6, submit

5.2. Route B Context Sources

Forward HyDE depends on a generation prompt, and the *context* supplied to that prompt is an open design choice. This section asks two questions: does injecting occupational-taxonomy context into the prompt improve forward HyDE over a job-description-only baseline, and if so, which context source is best? To answer these, we compare the variants foreshadowed in Section 4 — plain HyDE (no taxonomy), ESCO only, O*NET only, both combined, and a direct IDF-retrieval variant — against the occupation-anchored TaxHyDE used in the submitted system. Each variant changes only the B-context; all other routes and fusion settings are held fixed. Table 3 reports each variant twice: Route B in isolation and the best consensus-fusion MAP.

We first measure **plain HyDE** (job description only, no taxonomy in the prompt) against enriched variants. We then compare three Route B *context* modes for HyDE prompt enrichment: ESCO only, O*NET only, and both combined (dense retrieval over taxonomy records), distinct from occupation-anchored TaxHyDE used in the submitted system. We also evaluated a `faiss-idf` mode that bypasses occupation matching entirely, instead embedding individual JD segments and retrieving concepts directly from the full 44K-record taxonomy index ranked by IDF. This mode underperformed occupation-anchored TaxHyDE in weighted fusion on the development set (en-en: 0.8817 vs. 0.8869 for $A{+}B{+}D{+}E$;

Table 3
Forward HyDE MAP across taxonomy context sources, Route B in isolation and after consensus fusion (condition C_1, development set).

Setup: raw B-context (ESCO/O*NET/Both); fusion = consensus RRF (k=60).				
	en-en		**es-es**	
Context	Route B	Best fusion	Route B	Best fusion
None (plain HyDE)	0.8023	0.8808	0.8185	0.8869
ESCO	**0.8150**	**0.8848**	**0.8312**	**0.8907**
O*NET	0.7567	0.8808	0.7414	0.8799
Both	0.7548	0.8816	0.7957	0.8816

Plain HyDE: best fusion is $A{+}D{+}E$ (en), $B{+}D{+}E$ (es).
ESCO best fusion matches $B_{\text{esco}} + D + E$ (Table 4).

es-es: 0.8829 vs. 0.8832 for $B{+}D{+}E$) and is omitted from subsequent tables. Table 3 (condition C_1) shows Route B MAP in isolation and the best consensus-fusion MAP for each context source on the development set.

ESCO enrichment improves Route B in isolation by 0.013 MAP (en-en) and 0.013 MAP (es-es) relative to plain HyDE (0.8150 vs. 0.8023; 0.8312 vs. 0.8185). Under consensus RRF, ESCO enrichment gains $\approx$0.004 MAP over the best plain fusion per language (0.8848 vs. 0.8808 en-en; 0.8907 vs. 0.8869 es-es). Gains are modest but consistent on Route B. Once fused with Routes D and E, the lift shrinks because BM25 is the dominant individual route and the taxonomy signal is concentrated in Route B. Fusion winners differ by language ($A{+}D{+}E$ vs. $B{+}D{+}E$ for plain HyDE).

ESCO consistently outperforms O*NET and the combined source for forward HyDE under C_1 across both languages. We attribute this to a vocabulary alignment effect: ESCO skill labels (e.g., "failure analysis", "root cause analysis") match the terminology used in CVs, whereas O*NET task descriptions use job-description-style language (e.g., "Analyze defective products to determine root causes of failures"), creating a vocabulary mismatch when the generated profile is compared against CV embeddings. The combined source does not improve over ESCO alone, suggesting that O*NET's task-level specificity introduces noise rather than complementary signal. For the submitted system we instead use B^{anch} (occupation-anchored TaxHyDE: two-hop retrieval over the combined ESCO+O*NET index with IDF-ranked concepts; condition C_2), which aligns with Section 4 and outperforms raw O*NET-only B-context in this ablation.

5.3. Route Ablation

Table 4 (condition C_1) shows MAP for individual routes and key combinations on the development set. Routes A, D, and E are independent of B-context. Only B and C (which includes the forward HyDE profile) use B_{esco}.

Our best fused configurations exceed the official development baseline (0.8848–0.8907 vs. 0.704/0.645 en-en/es-es in Table 4). That baseline reaches Avg.MAP 0.67 on development but only 0.383 on test (Table 7), so the 10-query development split is easier than the 40-query test split for all systems.

Several observations emerge. First, BM25 (Route D) is the strongest individual route for both languages, outperforming all dense retrieval variants. This is counterintuitive given the general trend toward dense methods, but consistent with the finding that domain-specific terminology in job matching benefits from exact lexical matching [4].

Second, Reverse HyDE (Route E) is the second-strongest individual route for English, confirming that projecting CVs into job-description space provides a complementary signal to direct JD-CV comparison.

Third, multi-route fusion consistently outperforms any individual route. Under C_1, $B_{\text{esco}} + D + E$ achieves the highest MAP for both languages (0.8848 en-en, 0.8907 es-es). Adding all five routes degrades performance compared to the best three- or four-route sets, suggesting that redundant signals (A and C_{esco} overlap substantially with B_{esco} and E) can dilute fusion quality.

For weighted tuning and submission we switch to condition C_2 (B^{anch}, weighted RRF): English uses

Table 4
MAP of individual routes and multi-route fusions, with the official organizer baseline (condition C_1, development set). The organizer baseline corresponds to Codabench submission 582875.

Setup: B_{esco} context; fusion = consensus RRF (k=60).		
Configuration	**en-en**	**es-es**
Individual routes		
A (Dense)	0.8130	0.8290
B_{esco} (Forward HyDE)	0.8150	0.8312
C_{esco} (Concat)	0.8086	0.8312
D (BM25)	**0.8521**	**0.8474**
E (Reverse HyDE)	0.8422	0.8280
Multi-route fusion		
A+D	0.8683	0.8739
D+E	0.8779	0.8739
B_{esco}+D+E	**0.8848**	**0.8907**
C_{esco}+D+E	0.8806	0.8808
A+B_{esco}+D+E	0.8841	0.8858
A+B_{esco}+C_{esco}+D+E	0.8778	0.8797
Official baseline (TalentCLEF)	0.704	0.645

Organizer baseline: Avg.MAP 0.67; en-es MAP 0.582 (Codabench only).

Table 5
Per-route weight tuning under weighted RRF, top configurations versus equal weights (condition C_2, development set).

Setup: B^{anch}; weighted RRF (k=60); routes A, B, D, E. **en-en**					
	w_A	w_B	w_D	w_E	MAP
Equal wt.	1.0	1.0	1.0	1.0	0.8862
Best	1.0	1.5	1.5	1.5	**0.8869**
Rank 2	1.0	1.0	1.0	2.0	0.8866
Rank 4	0.5	1.0	1.0	1.0	0.8865

Setup: B^{anch}; weighted RRF (k=60); routes C, D, E. **es-es**				
	w_C	w_D	w_E	MAP
Equal wt.	1.0	1.0	1.0	0.8815
Best	1.0	1.5	1.0	**0.8832**
Rank 2	0.5	1.0	1.0	0.8828
Rank 3	1.0	1.5	1.5	0.8825

A+B^{anch}+D+E (retaining dense route A) and Spanish uses C^{anch}+D+E, where C^{anch} is Route C (Section 4.2.3) with the HyDE profile from occupation-anchored TaxHyDE. The job description and that profile are concatenated, embedded, and ranked against the CV corpus. Spanish does not fuse a separate Route B list because the forward HyDE text is already included in C^{anch}. This deliberately changes both B-context and route set relative to the C_1 winner (Table 5).

5.4. Weighted RRF

Under condition C_2, we tune per-route weights on the development set using a curated grid search over the submission route sets above (29 configurations for English, 17 for Spanish; k=60). Table 5 shows the top configurations alongside equal-weight (1.0) baselines.

Gains from weight tuning are marginal: +0.0007 for English and +0.0017 for Spanish over equal weights (from 0.8862 to 0.8869 and 0.8815 to 0.8832). Across the full curated grids (29 configurations for English, 17 for Spanish), MAP varied by approximately 0.0083 and 0.0077 respectively. Table 5 lists only the top ranks (spanning 0.0007 and 0.0017). This indicates the pipeline is robust to weight choices.

Table 6
Effect of SlideGar listwise reranking on fused rankings, with and without corpus-graph expansion (condition C_2, development set).

Lang	Graph	c	w	b	MAP	Δ
Setup: C_2; EN baseline MAP 0.8865 (rank 4 weights in Table 5).						
en	off	40	20	10	**0.8916**	+0.0051
en	on	40	20	10	0.8192	−0.0673
es	off	30	10	5	0.8428	−0.0404

Notably, both languages benefit from upweighting Route D (BM25), consistent with its strong individual performance. Route A (dense retrieval) benefits from equal or reduced weight on English, suggesting its signal is largely subsumed by the HyDE-based routes. These grids define our Codabench submission (condition C_2, with CLI label `occupation` for B^{anch}). After submission, further development reruns with weighted RRF (no SlideGar) found marginally higher MAP per language. English rose from 0.8869 to 0.8900 under an O*NET-filtered two-hop TaxHyDE variant (`occupation-onet`, same $A{+}B{+}D{+}E$ routes). Spanish rose from 0.8832 to 0.8934 under $B_{\text{esco}}{+}D{+}E$ (Table 3) instead of $C^{\text{anch}}{+}D{+}E$. We did not re-submit these settings. They show that B-context and route-set choice, not weight tuning alone, separated our submission from a stronger development upper bound.

5.5. Listwise Reranking

We evaluate SlideGar listwise reranking [22] with RankZephyr 7B on the C_2 weighted-RRF baselines from Table 5 (EN: $w_A{=}0.5$, $w_B{=}1.0$, $w_D{=}1.0$, $w_E{=}1.0$, MAP 0.8865, rank 4; ES: best weighted $C^{\text{anch}}{+}D{+}E$, MAP 0.8832), not the top English grid point (0.8869). Table 6 summarizes representative settings (c: budget, w: window, b: step; graph = corpus neighbor expansion).

Graph expansion consistently hurts both languages. Without it, reranking gains +0.0051 MAP on English only. Spanish reranking degrades in all settings, consistent with RankZephyr's English-only training (MS MARCO); translating the reranking window to English before judging relevance also failed (−0.0487 MAP). We therefore apply C_2 (B^{anch}, weighted RRF; Table 5) and SlideGar reranking ($c{=}40$, $w{=}20$, $b{=}10$, no graph expansion) for **en-en** only (development MAP 0.8916). For **es-es** we use weighted $C^{\text{anch}}{+}D{+}E$ without reranking (MAP 0.8832). Cross-lingual **en-es** uses Gemma-translated Spanish CVs with $B^{\text{anch}}{+}E$ and no SlideGar (see below).

5.6. Cross-Lingual Translation

For **en-es**, Spanish CVs are translated to English with **Gemma3:27B** so BM25 and dense routes share a language space. Sentence-level NLLB-200 [23] was rejected after qualitative checks: it corrupted section headers, company names, and dates, which harms lexical and structural matching. Gemma document-level translation preserved CV layout but was not fully ablated before submission. MAP for **en-es** is therefore reported from Codabench test runs only.

5.7. Test Set Results

Table 7 shows our test set performance on the official TalentCLEF 2026 Task A benchmark (best submission per team).

On the official leaderboard, CYUT ranks **12th among 33 entrants** by best submission—using team names where listed and otherwise the participant login (also **25th of 94** individual valid test submissions; submission ID 702388). Our average MAP of 0.609 remains substantially above the organizer baseline (0.383) but below the leading systems.

The most notable observation is the large generalization gap between development and test performance. The organizer baseline shows the same pattern: Avg.MAP falls from 0.67 on development (Table 4) to 0.383 on test. Our **final** development MAP on the assembled pipeline reaches **0.8916** (en-en,

Table 7
Official test-set leaderboard standings for CYUT against the top entrants and organizer baseline (official benchmark). One row per entrant (team name, or participant login when no team name is listed), keeping each entrant's best submission by AVG_MAP(en-en, es-es); ranks #1–#3 are leaders, CYUT and the organizer baseline are shown for context. Avg.MAP = mean of MAP(en-en) and MAP(es-es).

Team	Avg.MAP	en-en	es-es	en-es
classum (#1)	0.714	0.712	0.716	0.704
QTPride (#2)	0.663	0.662	0.664	0.636
bipboopbipboop (#3)	0.653	0.655	0.651	0.644
CYUT (#12)	0.609	0.597	0.621	0.585
Baseline (#29)	0.383	0.400	0.366	0.356

with SlideGar) and **0.8832** (es-es, weighted fusion only), yet test MAP falls to 0.597 (en-en) and 0.621 (es-es). A drop of 0.2946 points on English and 0.2622 on Spanish.

We identify three likely contributing factors. First, the test set is likely more challenging, with weaker query–document links and less-relevant context for HyDE generation. Second, with only 10 development queries per language and tuning across B-context, fusion weights, and reranking parameters, some of the development–test gap may reflect overfitting to the small labeled split rather than architecture alone. Third, cross-lingual performance (en-es MAP 0.585) remains our weakest mode relative to monolingual tracks. The winning team achieves near-parity across all three modes (0.704–0.716 MAP), suggesting a more language-agnostic architecture.

Codabench MRR for submission 702388 remains competitive: MRR(en-en) = 0.817 and MRR(es-es) = 0.859, indicating that the system places a relevant candidate at rank 1. We did not compute MRR locally on the development set (10 queries per language).

6. Conclusion and Future Work

Our participation in TalentCLEF 2026 Task A showed that a training-free multi-view fusion system substantially outperforms the official organizer baseline on multilingual job-person matching without task-specific labels (12th of 33 entrants by best submission, 25th of 94 individual submissions, and average MAP 0.609 vs. test baseline 0.383 and development baseline 0.67), while remaining below the strongest tuned systems. Weighted RRF effectively combines dense, sparse, and HyDE-based routes. Listwise reranking helped English only under our settings.

Taxonomy-grounded forward HyDE is a deliberate, distinguishing component of this study. Under condition C_1 (raw ESCO B-context, consensus RRF), ESCO enrichment outperforms plain HyDE (job-only prompts) with a consistent lift on Route B in isolation ($\approx$0.013 MAP, both languages) and a smaller gain on the best fused configurations ($\approx$0.004 MAP per language). Our Codabench submission instead uses occupation-anchored TaxHyDE with weighted RRF (C_2). The attenuation under fusion is itself informative: because BM25 is the strongest individual route and occupational taxonomy primarily sharpens the HyDE generation route, fusing with a strong lexical view caps how much taxonomy enrichment can move the final ranking.

Future work should test whether taxonomy signals survive fusion when lexical routes are weaker or when enrichment is applied to reverse HyDE and other routes. Fusion and B-context choices on larger labeled splits to reduce development set overfitting, and to pursue more robust cross-lingual matching without English-only rerankers or full-corpus translation can also be evaluated. We did not complete a gender-bias analysis in this submission. Future work should evaluate our rankings on the Task A bias-controlled track (RBO).

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 4.6 Extended in order to: Grammar and spelling check. Furthermore, the authors used Cursor Auto to help write the document in LaTeX format to help with formatting. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[3] J. Shen, Y. Juan, S. Zhang, P. Liu, W. Pu, S. Vasudevan, Q. Song, F. Borisyuk, K. Q. Shen, H. Wei, Y. Ren, Y. S. Chiou, S. Kuang, Y. Yin, B. Zheng, M. Wu, S. Gharghabi, X. Wang, H. Xue, Q. Guo, D. Hewlett, L. Simon, L. Hong, W. Zhang, Learning to retrieve for job matching, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2024.

[4] N. Tonellotto, C. Macdonald, Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3459637.3482159.

[5] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2023.

[6] I. Mackie, J. Dalton, Bridging the question–answer gap in retrieval-augmented generation: Hypothetical prompt embeddings, IEEE Access (2025). doi:10.1109/ACCESS.2025.3573053.

[7] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2009. doi:10.1145/1571941.1572114.

[8] C. Zhu, H. Zhu, H. Xiong, C. Ma, F. Xie, P. Ding, P. Li, Person-job fit: Adapting the right talent for the right job with joint representation learning, ACM Transactions on Management Information Systems 9 (2018). doi:10.1145/3234465.

[9] C. Yang, Y. Hou, Y. Song, T. Zhang, J.-R. Wen, W. X. Zhao, Modeling two-way selection preference for person-job fit, in: Proceedings of the 16th ACM Conference on Recommender Systems, Association for Computing Machinery, 2022. doi:10.1145/3523227.3546752.

[10] X. Yu, R. Xu, J. Zhang, Z. Yu, ConFit: Improving resume-job matching using data augmentation and contrastive learning, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2024.

[11] X. Yu, R. Xu, C. Xue, J. Zhang, Z. Yu, ConFit V2: Improving resume-job matching using hypothetical resume embedding and runner-up hard-negative mining, arXiv preprint arXiv:2502.12361 (2025).

[12] I. A. Kostis, et al., Towards an integrated retrieval system to semantically match CVs, job descriptions and curricula, in: 26th Pan-Hellenic Conference on Informatics, Association for Computing Machinery, 2022. doi:10.1145/3575879.3575985.

[13] S. Vaishampayan, H. Leary, Y. B. Alebachew, L. Hickman, B. Stevenor, W. Beck, C. Brown, Human and LLM-based resume matching: An observational study, in: Findings of the Association for

Computational Linguistics: NAACL 2025, Association for Computational Linguistics, 2025, pp. 4808–4823.

[14] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[15] 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, Zenodo, version 0.3.0, 2026. URL: https://zenodo.org/records/19652670. doi:10.5281/zenodo.19652670.

[16] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 (2024).

[17] Gemma Team, Gemma 3, arXiv preprint arXiv:2503.19786 (2025).

[18] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The Faiss library (2024). arXiv:2401.08281.

[19] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[20] L. Liu, M. Zhang, Exp4Fuse: A rank fusion framework for enhanced sparse retrieval using large language model-based query expansion, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 163–173. doi:10.18653/v1/2025.findings-acl.9.

[21] R. Pradeep, S. Sharifymoghaddam, J. Lin, RankZephyr: Effective and robust zero-shot listwise reranking is a breeze!, arXiv preprint arXiv:2312.02724 (2024).

[22] M. Rathee, S. Chaudhary, F. Nargesian, F. Diaz, Guiding retrieval using LLM-based listwise rankers, arXiv preprint arXiv:2501.09186 (2025).

[23] NLLB Team, No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).