Taxonomy-Aware Hybrid Retrieval with Bounded LLM Reranking for TalentCLEF 2026 Task A

Arij Riabi[1,*], Malek Essayeh[1]

[1]Devoteam

Abstract

This paper describes our submission to TalentCLEF 2026 Task A, contextualized job-person matching. The task requires ranking candidate CVs for job offers in English, Spanish, and English-to-Spanish cross-lingual settings. We use taxonomy-normalized retrieval as the backbone for multilingual job-person matching, and apply constrained large language model (LLM) reranking to improve top-rank precision when grounded in European Skills, Competences, Qualifications and Occupations (ESCO) evidence rather than used as an unconstrained full-corpus ranker. The submitted system combines hybrid retrieval, ESCO-normalized skill evidence, graph calibration, and bounded listwise LLM refinement. As a procedural safeguard, demographic cues are masked before scoring. The final submitted system improved over a strong retrieval-only graph baseline, with the largest gain observed in the English-to-Spanish direction. Overall, the results suggest that a strong retrieval backbone makes bounded LLM reranking safer, with the largest benefit in the cross-lingual setting, and that ESCO graph evidence works best in our system as a calibration signal within that retrieval-first design.

Keywords

Talent matching, Information retrieval, Skill extraction, ESCO, Large language models, Multilingual retrieval, Demographic masking

1. Introduction

Talent matching is a practical information retrieval problem with important scientific and societal implications. TalentCLEF studies skill and job-title intelligence for human capital management [1, 2]. In TalentCLEF 2026 Task A, contextualized job-person matching, systems rank candidate CVs for job offers in monolingual English, monolingual Spanish, and English-to-Spanish cross-lingual settings [3]. Unlike generic document retrieval, job-person matching requires evidence that is both semantic and occupational: relevant CVs must satisfy hard skills and qualifications, but also match seniority, domain context, role family, and language requirements. At the same time, rankings should avoid relying on demographic surface cues that are not job-relevant.

This setting creates a practical ranking trade-off. LLMs can compare candidates in a nuanced way, but using them as full-corpus rankers is expensive, difficult to reproduce, and vulnerable to formatting errors. Embedding and lexical retrievers are faster and more stable, but they often overreward broad professional vocabulary such as management, communication, or quality even when the occupational domain is wrong. We address this trade-off with a staged ranking pipeline: retrieval first identifies a plausible candidate pool, ESCO evidence calibrates occupational relatedness, and the LLM is reserved for top-rank tie-breaking and qualitative comparison.

We treat the submission as an empirical design study around three questions. First, can a multilingual retrieval stack produce a competitive full ranking without supervised fine-tuning on TalentCLEF relevance labels? Second, can ESCO normalization and graph structure reduce false positives caused by generic skill overlap? Third, does a listwise LLM reranker add value when its input is restricted to the strongest retrieved candidates and its prompt is grounded in taxonomy-normalized evidence?

The system combines standard reusable components with task-specific orchestration. BM25, Reciprocal Rank Fusion (RRF), BGE-M3, SkillMatch MPNet, ESCOXLM-R, Node2Vec/Word2Vec, and the LLM deployment are used as off-the-shelf building blocks. The contribution lies in how these

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

*Corresponding author.

arij.riabi@devoteam.com (A. Riabi)

blocks are arranged for TalentCLEF Task A: a retrieval-first backbone, ESCO-normalized evidence for occupational calibration, and bounded LLM reranking over the strongest retrieved candidates. The two most task-specific choices are the development-corpus reweighting of ESCO graph edges with inverse document frequency (IDF) and hub-node penalties, and ESCO-normalized adaptive prompt selection for the listwise LLM reranker. The paper makes three contributions aligned with the research questions:

- we build a strong retrieval backbone by combining lexical, multilingual dense, and skill-specialized retrieval into complete job-person rankings;
- we add taxonomy-normalized calibration through extracted and mined ESCO skill evidence, semantic skill coverage, and a development-weighted ESCO graph;
- we use adaptive, ESCO-grounded prompt selection for bounded listwise LLM reranking, while preserving the retrieval-only ranking as the fallback when LLM output is incomplete or noisy.

The final system outperformed our retrieval-only graph baseline on the official test results. The results suggest that retrieval should remain the primary ranking layer, with ESCO evidence and LLM reranking used as bounded adjustments over strong retrieved candidates.

2. Related Work

Prior work on HR-NLP frames job-person matching as more than generic document similarity. Surveys of human-resource NLP and computational job-market analysis highlight domain terminology, standardized taxonomies, multilingual variation, sparse supervision, and fairness risks as recurring constraints [4, 5]. These constraints motivate our retrieval-first setting: the system must produce complete rankings, but the evidence used to order candidates should remain inspectable and grounded in job-relevant signals.

Taxonomy-normalized skill representation is the first relevant line of work. SkillSpan, ESCOXLM-R, ESCO entity linking, nearest-neighbor occupational skill extraction, and zero-shot or LLM-assisted ESCO matching all show different ways to map noisy job-market language into a shared skill inventory [6, 7, 8, 9, 10, 11, 12]. Related work on multilingual occupational knowledge graphs and data-driven taxonomy construction shows that occupational structure can add information that is missing from raw text overlap [13, 14]. In our system, these ideas appear as ESCO spans, mined ESCO candidates, and graph-based calibration signals rather than as a hard taxonomy filter.

The second relevant line of work concerns ranking architecture. Recent job and talent matching systems use role-aware retrieval, contrastive training, hard-negative mining, LLM-generated job or candidate representations, and prompt-based matching [15, 16, 17, 18]. Given the absence of Task A training labels, our system combines reusable retrieval signals and reserves the LLM for bounded reranking. This follows the broader evidence that listwise LLM prompting can improve document reranking [19, 20], including in zero-shot cross-lingual retrieval [21], while keeping the LLM anchored to a complete retrieval ranking.

Finally, the hiring context constrains how reranking evidence should be used. LLM job-resume matching can vary with demographic attributes, and cross-lingual job retrieval can vary with the language of the query or candidate profile [22, 23]. This motivates demographic masking before scoring, prompts focused on job-related evidence, and separate reporting for the English-to-Spanish condition.

3. Task and Data

TalentCLEF Task A frames job-person matching as an ad-hoc ranking task [3]. We used the official TalentCLEF 2026 corpus [24]. Each job offer is treated as a query and each candidate CV is treated as a document. The system must produce TREC-style ranked lists with one ordering per query:

```
<query_id> Q0 <doc_id> <rank> <score> <run_tag>
```

We evaluate three directions: `en-en`, where English job offers are ranked against English CVs, and `es-es`, where Spanish job offers are ranked against Spanish CVs. We also evaluate the cross-lingual `en-es` setting, where English job offers are ranked against Spanish CVs.

The official data layout contains separate `queries` and `corpus` folders for English and Spanish. No training set is provided. The development split contains relevance judgments for local validation, while the test labels are hidden. The development split is small, so we use it to compare design choices rather than to train a supervised neural ranker. This motivates a configuration based on reusable signals—pretrained retrieval models, ESCO normalization, graph calibration, and constrained reranking—instead of fitting a model to Task A relevance labels.

The official evaluation uses ranking metrics including Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), precision at cutoff ranks (P@5, P@10, and P@100), Normalized Discounted Cumulative Gain (NDCG), and a bias-oriented ranking comparison metric based on Rank Biased Overlap (RBO) [25]. In the controlled-bias benchmark, the organizer-reported average RBO uses only the `en-es` and `es-es` directions because these are the directions with relevant gender-related annotations. We emphasize MAP in the analysis because it was the main metric used to compare systems in the benchmark categories and because it rewards ranking relevant CVs early while still evaluating the full ordering. We used the development split for local parameter selection and the hidden test split only for final Codabench submissions.

4. System Overview

Figure 1 summarizes the submitted pipeline. Although the implementation uses several signals, the data flow is simple: rank the full CV corpus, calibrate the ranking with ESCO-normalized skill evidence, and apply listwise LLM reranking only to the most plausible candidates.

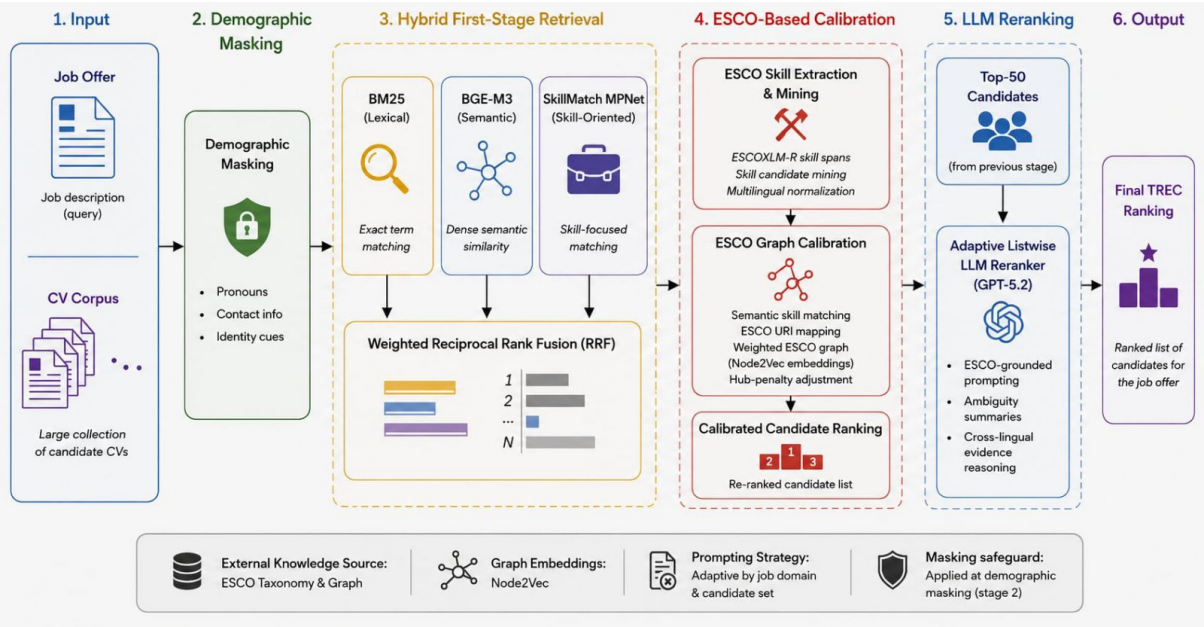

Figure 1: Submitted retrieval-first talent matching pipeline.

Table 1 lists the concrete resources behind the pipeline. This heterogeneous mix uses classical retrieval and graph tools for robust scoring, while newer multilingual and LLM components handle semantic matching and top-candidate comparison.

Table 2 reports the final values needed to reproduce the submitted configuration and to interpret the equations below. We include the ranking, skill, graph, and LLM fusion parameters rather than every implementation flag.

Table 1
Main resources used by the submitted system.

Component	Model or resource	Role
Lexical retrieval	BM25-style scorer	Exact terminology
Dense retrieval	BAAI/bge-m3	Multilingual similarity
Skill retrieval	SkillMatch MPNet	Labor-market skill similarity
Skill extraction	jjzha/escoxlmr_skill_extraction	Skill spans
Skill graph	ESCO + Node2Vec	Related-skill calibration
LLM reranker	Azure OpenAI GPT-5.2 deployment	Top-50 listwise refinement

Table 2
Final submitted configuration.

Component	Selected values
RRF retrieval fusion	$k = 60$; $(w_{\mathrm{BM25}}, w_{\mathrm{BGE}}, w_{\mathrm{SkillMatch}}) = (0.8, 0.8, 0.6)$; normalized component-score coefficient 0.05
Semantic skill matching	semantic/exact/ESCO-URI weights $(0.65, 0.20, 0.15)$; semantic match threshold $\tau_s = 0.55$; ESCO mining threshold 0.62
ESCO graph calibration	graph fusion weight $\lambda_g = 0.18$; ESCO mapping threshold 0.65; minimum graph score 0.55; low-score penalty multiplier 0.75
Listwise LLM reranking	top-50 candidates; 1800 characters per CV; LLM fusion weight $\lambda_l = 0.60$; ESCO-normalized adaptive prompting and ambiguity summary enabled
Demographic masking	basic contact/demographic masking applied to all scoring text

As shown in Figure 1, demographic masking is applied before all scoring components.

4.1. Hybrid First-Stage Retrieval

The first-stage ranker is an ensemble of three full-corpus retrieval channels. Each channel addresses a different failure mode. The lexical channel uses BM25-style scoring and captures exact requirement wording, which is important for certifications, named tools, and explicit domain terms. The dense channel uses BGE-M3 [26], a multilingual embedding model designed for multi-function and long-context retrieval; it provides the main cross-lingual semantic signal. The skill-specialized channel uses the SkillMatch MPNet curriculum retriever [27], a model trained for job-skill matching; it gives more weight to labor-market skill semantics than to generic textual similarity. We kept both BGE-M3 and SkillMatch MPNet because they serve different purposes: BGE-M3 provides broad multilingual document similarity, whereas SkillMatch MPNet is a narrower labor-market signal that is informative when the decisive evidence is a skill expression rather than the overall document topic.

The three ranked lists are fused with weighted RRF [28]. For a document d, each component ranker contributes a score inversely proportional to its rank, plus a small normalized component score term:

$$S_{\mathrm{rrf}}(d) = \sum_i w_i \left(\frac{1}{k + r_i(d)} + 0.05\, \hat{s}_i(d) \right), \tag{1}$$

where $r_i(d)$ is the rank assigned by component i, $\hat{s}_i(d)$ is the component score normalized by the maximum score in that list, w_i is the component weight, and k is the RRF smoothing constant. The fused scores are then normalized before later calibration. The component weights were selected on the development set to keep lexical and multilingual dense evidence strong while retaining the skill-specialized signal. BGE-M3 and SkillMatch embeddings are computed locally and cached, which made repeated development runs reproducible.

4.2. Skill Extraction and Semantic Skill Matching

Skill spans are extracted with the multilingual ESCOXLM-R skill extraction model. This component is different from the SkillMatch retriever: ESCOXLM-R identifies candidate skill mentions in the text, while SkillMatch MPNet ranks CVs according to skill-oriented semantic similarity. ESCO skill mining is a third operation: it retrieves taxonomy labels from text chunks to recover likely skills that were not returned by the token-level extractor. Thus, extraction finds spans, skill retrieval ranks documents, and ESCO mining adds normalized taxonomy candidates. Extracted job and CV skill spans are then compared semantically, rather than only through exact string matching. This is important for Spanish and cross-lingual matching, where equivalent skills may appear with different surface forms.

Let J be the extracted or mined job skill spans and C the corresponding CV skill spans. For each job skill, the system finds the most similar CV skill in embedding space and applies a thresholded similarity function ϕ. The semantic score is:

$$\begin{aligned} S_{\text{skill}} = 0.65 \frac{1}{|J|} \sum_{j \in J} \max_{c \in C} \phi(\cos(e_j, e_c)) \\ + 0.20 \frac{|J \cap C|}{|J|} \\ + 0.15 \frac{|U_J \cap U_C|}{|U_J|}, \end{aligned} \tag{2}$$

where e_j and e_c are BGE-M3 span embeddings, and U_J and U_C are the ESCO URI sets linked from job and CV skills. The function ϕ converts raw cosine similarity into usable evidence by suppressing weak matches:

$$\phi(x) = \begin{cases} 0, & x < \tau_s, \\ x, & x \geq \tau_s, \end{cases}$$

where τ_s is a development-selected semantic-match threshold. Thus, a nearby CV skill contributes to the score only when it is sufficiently close to a job skill in the multilingual embedding space.

The skill score mainly rewards semantically similar skills, keeps exact matches for precise requirements, and adds extra confidence when both sides map to the same ESCO skill concept.

We also enable ESCO skill mining. This component embeds short text chunks and n-grams with BGE-M3, retrieves nearby multilingual ESCO skill labels, and keeps high-confidence candidates. It is used as a recall booster when the token classifier misses implicit, unusual, or long-form skill descriptions. The mining threshold is intentionally conservative: mined skills should recover missing evidence without introducing many weakly related concepts.

4.3. ESCO Skill Graph Calibration

The submitted system includes a URI-based multilingual ESCO skill graph built from the ESCO taxonomy [29]. The motivation is that job-CV fit is often relational: two candidates may not share the exact same extracted skill labels, yet their skills may be close in the ESCO occupation-skill structure. The graph therefore acts as a smoothing mechanism over normalized skill concepts.

The graph source is the official ESCO v1.2.0 RDF export. We extract English and Spanish preferred and alternative labels from the RDF file, attach them to the same ESCO URIs, and then build the weighted multilingual graph used by the submitted run. The final graph contains 16,978 nodes: 13,939 skills and 3,039 occupations. Its 132,511 edges are mostly occupation-skill relations, with 126,051 occupation-skill edges and 6,460 hierarchy or skill-skill edges.

The edge weighting is intended to make specialized skills more influential than broad generic ones. First, Task A development documents are mapped to ESCO skill URIs. ESCOXLM-R extracts candidate skill phrases, BGE-M3 embeds those phrases and the multilingual ESCO labels, and a phrase is assigned to its nearest ESCO skill label when the cosine similarity exceeds the graph-construction mapping

threshold, set to 0.58 for the graph-building step. From these mapped document frequencies we compute a capped IDF-style weight,

$$\begin{aligned} \mathrm{idf}_{\mathrm{raw}}(u) &= \log\left(\frac{N+1}{\mathrm{df}(u)+1}\right)+1, \\ \mathrm{idf}(u) &= \min\left(4.0, \max\left(1.0, \mathrm{idf}_{\mathrm{raw}}(u)\right)\right), \end{aligned} \tag{3}$$

where N is the number of development documents with at least one mapped skill. For an edge (u, v), the score uses the larger endpoint IDF and a degree-based hub penalty $h(u, v)$:

$$h(u,v) = 1 + 0.35\,\frac{\log\left(1+\max(\deg(u),\deg(v))\right)}{\log\left(1+\deg_{\max}\right)}. \tag{4}$$

$$w(u,v) = \frac{b(u,v)\max(\mathrm{idf}(u),\mathrm{idf}(v))}{h(u,v)}. \tag{5}$$

The base weight $b(u, v)$ is 1.0 for occupation-skill edges and 0.35 for hierarchy or skill-skill edges. This hub penalty reduces the effect of highly connected generic concepts such as communication, management, quality, or teamwork. Without this calibration, generic skills can produce high graph similarity between candidates who are generally employable but occupationally mismatched.

These graph-construction constants were fixed before the final ranking sweeps: the IDF cap prevents rare or unmapped concepts from dominating, the occupation-skill base weight keeps direct occupational evidence strongest, and the lower hierarchy/skill-skill base weight treats related skills as softer evidence.

We then train Node2Vec/Word2Vec embeddings from biased random walks over the weighted graph [30, 31]. The submitted model uses 128 dimensions, 160 walks per node, walk length 30, context window 10, and Node2Vec parameters $p = 0.5$ and $q = 2.0$. We do not apply a hard shortest-path hop cutoff at inference time; graph locality is instead encoded by the weighted random-walk neighborhoods seen by Node2Vec.

At inference, extracted and mined job/CV skill phrases are mapped to ESCO labels with BGE-M3 nearest-neighbor matching. The ranking-time mapping threshold is 0.65, so only high-confidence label matches become ESCO URIs. The URI Node2Vec vectors are averaged into job and CV centroids, and centroid cosine similarity becomes a graph score. The graph score calibrates the retrieval score; it does not replace the retriever.

$$S_{\mathrm{graph}}(q,d) = \cos\left(\frac{1}{|U_q|}\sum_{u\in U_q} z_u, \frac{1}{|U_d|}\sum_{u\in U_d} z_u\right), \tag{6}$$

where z_u is the Node2Vec representation of ESCO URI u. The calibrated retrieval score is:

$$S_{\mathrm{cal}}(q,d) = (1-\lambda_g)S_{\mathrm{ret}}(q,d) + \lambda_g S_{\mathrm{graph}}(q,d), \tag{7}$$

where $\lambda_g = 0.18$ is the graph calibration weight: small values keep the retriever dominant, while larger values give more influence to ESCO graph relatedness. The final run uses a minimum graph score of 0.55 and a low-score multiplier of 0.75; the multiplier is applied only when a graph score exists and falls below the threshold. This is why the graph acts as a calibration layer rather than a hard filter. We selected the graph weight, ranking-time mapping threshold, minimum score, and penalty multiplier on the development split through local ablations, using the criterion that taxonomy evidence should move close retrieval ties without allowing graph similarity to override the retrieval backbone.

4.4. Adaptive Listwise LLM Reranking

The LLM is used only after calibrated retrieval. For each job and language direction, the system takes the top 50 candidates from the calibrated retrieval ranking and sends them in one listwise batch. Each candidate is represented as a `<doc id="...">` block, not as an independent pairwise comparison.

Candidate text is demographically masked and truncated to 1,800 characters using a head/tail split, so the beginning and end of long CVs remain visible. We used an Azure OpenAI GPT-5.2 deployment configured with temperature 0.0 and a 16,384-token completion budget. The same English instruction template is used in all directions; job and CV text remain in their original language.

The prompt has four parts: an HR ranking role instruction, optional adaptive focus instructions, the masked job description plus the 50 candidate document blocks, and strict output rules requiring TREC run lines. In the submitted configuration, adaptive prompting is enabled. The inserted focus block is built from structured job requirements, selected prompt-pool entries, and retriever-derived ambiguity signals. This makes the prompt query-specific while keeping the output format fixed.

For the final combined order of m candidates, the final score is:

$$S_{\text{final}}(q,d) = 100\left[(1-\lambda_l)\hat{S}_{\text{ret}}(q,d) + \lambda_l \frac{m - r_o(d) + 1}{m}\right], \tag{8}$$

where $r_o(d)$ is the rank in the combined order: parsed LLM candidates first, then any unparsed top-50 candidates in retrieval order, then the remaining retrieval results. The value λ_l controls the influence of the listwise ordering. We set $\lambda_l = 0.60$ so the LLM can adjust top-rank order while the retrieval score remains an anchor.

Listing 1: Submitted listwise reranking prompt template.

```
You are an expert HR recruiter and retrieval ranking judge.
Rank candidate CV documents for one job description.

The job can belong to any profession or sector. Judge relevance only from
job-related evidence in each CV.

Adaptive Focus Instructions:
{adaptive_focus_instructions}

Ranking priorities:
1. Same or very closely related occupation/domain.
2. Mandatory skills, tasks, tools, licenses, certifications, and qualifications.
3. Direct evidence of relevant work experience and seniority.
4. Domain, language, education, mobility, or schedule requirements when explicit.
5. Soft skills only after core requirements are satisfied.

Hard rules:
- If a CV is from an unrelated occupation/domain, rank it below related CVs even
  when generic words overlap.
- Do not reward demographic details, names, pronouns, or contact information.
- For English-Spanish matching, accept equivalent evidence across languages.
- Output every provided candidate exactly once.
- Do not invent document ids.

Query ID: {query_id}
Run Tag: {run_tag}

Job Description:
{job_text}

Candidate Documents:
{candidates_text}

Output ONLY TREC run lines in this exact format:
<q_id> Q0 <doc_id> <rank> <score> <run_tag>

Rules:
- exactly six space-separated columns per line
- second column must be Q0
- q_id must be {query_id}
- run_tag must be {run_tag}
- rank starts at 1
- scores must be numeric between 0 and 1
- no explanations, no JSON, no markdown, no headers
```

The structured job requirements are not generated by an LLM. They are produced by deterministic rules over the job text before the LLM call. The extractor tokenizes the job and fills fixed fields for required and preferred skills, seniority, certifications, languages, management, education, tools or equipment, domain, and location or mobility. These fields come from small English/Spanish keyword vocabularies and marker rules; for example, seniority is inferred from terms such as “senior”, “lead”, “junior”, or “entry level”, while preferred skills are separated when the job contains markers such as “preferred”, “nice to have”, “plus”, “deseable”, or “valorable”.

The ESCO-normalized job profile is a combination of this rule-based requirement summary, the original job text, and ESCOXLM-R/ESCO-mined skill labels and URIs. It is used to select relevant prompt-pool instructions through lexical similarity and focus-specific boosts; it does not ask another

model to extract requirements. The prompt template then simply inserts the selected fields and instructions into the fixed listwise reranking prompt. Thus, the adaptive focus block contains: (i) rule-based structured job requirements, (ii) selected prompt-pool instructions, and (iii) a retriever ambiguity summary covering discriminative evidence, common generic overlap, hard constraints, and related-domain false-positive risks. Listing 2 shows a shortened example of the inserted block. The exact fields and selected instructions vary by query, but the headings and bullet types follow the submitted implementation.

Listing 2: Shortened example of adaptive focus instructions inserted into the listwise prompt.

```
Adaptive Focus Instructions:
Structured job requirements extracted for this query:
required_skills: distribution, inventory, logistics, shipping, warehouse
certifications: license
languages: english, spanish
tools_or_equipment: forklift, scanner, warehouse
domain: logistics, manufacturing
location_or_mobility: shift

Selected adaptive focus instructions from the prompt pool:
1. [operations_logistics] For operations, logistics, distribution,
   manufacturing, quality, and warehouse roles, emphasize process execution,
   inventory, receiving/shipping, safety, quality control, root-cause analysis,
   and productivity.
2. [certification] Pay special attention to required or preferred
   certifications, professional licenses, safety permits, regulated
   credentials, and formal qualifications.
3. [cross_lingual] For cross-lingual matching, recognize English-Spanish
   equivalents for titles, skills, tools, sectors, certifications, and
   responsibilities.

Retriever ambiguity summary:
- Job-specific evidence to discriminate: forklift, inventory, shipping
- Common top-candidate overlap, treat as weak unless supported by domain
  evidence: communication, management, quality
- Hard constraints to verify: certifications: license; languages: english,
  spanish; location_or_mobility: shift
- False-positive risk: generic management, quality, communication, or project
  experience from unrelated domains.
```

Parsing accepts only valid six-column TREC lines for the expected query, run tag, and candidate ids. Lines with the wrong query id, wrong run tag, invalid document id, invalid rank, invalid score, or extra/missing columns are ignored. If the same candidate appears more than once, the parser keeps the earliest rank. Valid parsed candidates are kept in LLM order, duplicate ids are deduplicated, and missing candidates are appended by calibrated retrieval rank, preserving a complete ranking when the LLM output is incomplete. If the LLM request fails, the connector returns a non-TREC error payload; this parses as zero valid listwise lines, so the ranking falls back to the calibrated retrieval order.

The reranker is therefore bounded in three ways. Candidate cost is bounded because the LLM never sees more than 50 CVs per job and never scores the full corpus. Token cost is bounded by the 1,800-character CV limit and the 16,384-token completion cap. Call cost is bounded because listwise mode makes one LLM request per query-direction rather than one request per candidate or candidate pair; in the official test configuration, this corresponds to 120 listwise calls across `en-en`, `es-es`, and `en-es`.

5. Results

This section first reports development diagnostics used during system tuning, then reports the official test results for the submitted systems.

Table 3 summarizes staged development ablations used as diagnostic/tuning evidence. The table suggests which cumulative changes were useful on the development split: hybrid retrieval over the BM25-style baseline, soft ESCO graph calibration over the retrieval backbone, and bounded listwise LLM reranking over the calibrated candidate pool. The comparison is cumulative rather than fully factorial, so it should not be read as independent evidence for each component. It does not separately isolate pure BM25, each dense retriever, ESCO mining, adaptive prompt selection, or the retriever ambiguity summary.

The development scores are useful as tuning diagnostics. In these local runs, hybrid retrieval improves over the BM25-style baseline, soft ESCO graph calibration raises the average MAP mainly through the

Table 3
Development-set ablations in MAP (%) and what each staged variant represents.

System	en-en (%)	es-es (%)	Avg. (%)	What it represents
BM25-style ranker baseline	80.04	76.82	78.43	Lexical/BM25-centered baseline with lightweight HR heuristics
Hybrid retrieval backbone	83.46	79.77	81.61	RRF fusion of BM25, BGE-M3, and SkillMatch before graph or LLM calibration
Hybrid + soft ESCO graph	83.01	82.85	82.93	Taxonomy-normalized skill and graph calibration without LLM reranking
Hybrid + soft ESCO graph + LLM	84.60	84.52	84.56	Top-50 listwise refinement over the calibrated retrieval pool
Hybrid + hard ESCO role gate	82.48	80.17	81.33	Test of a stricter occupational taxonomy penalty

Spanish ranking, and the LLM variant produces the best development result. On development, bounded listwise LLM reranking adds 1.63 points over the graph-only system; a paired query-level bootstrap over the 20 monolingual development queries gives a 95% confidence interval of +0.56 to +3.05 points. The stricter role-gate row is kept as a diagnostic contrast: in tuning, ESCO evidence was more useful when it adjusted retrieved candidates rather than excluding them. Because each monolingual development split contains only 10 queries, these results are best interpreted as directional ablations; the official test results below are used to check whether the same pattern transfers.

Table 4 reports the organizer-provided Task A benchmark MAP values (%) for our two DevoMatcher submissions. These hidden-test comparisons are between our two submitted variants only, not against the full Task A leaderboard. Official test results show a similar pattern: the retrieval-only variant remains competitive, but the bounded LLM system improves both the monolingual average and the cross-lingual `en-es` direction. On the official test set, the LLM system improves the monolingual average by 2.27 points. The largest absolute gain is obtained for `en-es`, where MAP increases by 4.29 points. These gains indicate that LLM reranking has the smallest effect where surface matching already orders candidates effectively, and the largest effect in `en-es`, where English job evidence and Spanish CV evidence must be aligned semantically.

Table 4
Task A benchmark MAP (%) for the submitted variants.

System	en-en (%)	es-es (%)	Avg. (%)	en-es (%)
Retrieval + graph, no LLM	58.77	56.47	57.62	48.64
Final bounded LLM system	59.66	60.12	59.89	52.93

In Table 4, `Avg.` averages only the `en-en` and `es-es` monolingual directions. The `en-es` column is reported separately as the cross-lingual condition.

Controlled-bias ranking stability is discussed separately in Section 6.

5.1. Qualitative Development Examples

The development diagnostics illustrate both the benefit and the risk of taxonomy-based semantic matching. A positive case appears for query `36044`, a Failure Analysis Engineer position. Relevant CV `14823` did not use the exact same title, but described reliability engineering, Failure Mode and Effects Analysis, process failures, root cause analysis, corrective actions, and process control. This candidate moved from rank 66 in an earlier multi-fused configuration to rank 8 in the graph-enhanced

configuration. Its lexical score remained moderate (52%), while its macro-skill score increased to 97%. This illustrates the intended behavior of the graph layer: ESCO-normalized evidence links different phrasings of related technical skills without requiring exact token overlap.

The same mechanism can also overestimate related but non-relevant occupations. For query `46795`, a Paralegal role focused on privacy and information management, CV `9130` was ranked second but was not marked relevant in the development qrels. The CV contained strong evidence for privacy governance, regulatory compliance, legal research, and risk assessment, leading to high embedding and macro-skill scores. However, its experience profile corresponded to a senior compliance and ethics manager rather than a paralegal support role. This example illustrates why ambiguity summaries and LLM reranking are useful: taxonomy similarity can surface related evidence, but role level and function still need to be judged downstream.

5.2. Error Analysis

Development diagnostics and submitted-run logs point to four recurring error categories. First, skill extraction is occasionally missing or noisy. Even with ESCOXLM-R and ESCO skill mining, some relevant evidence appears in long experience descriptions, project narratives, or abbreviated tool names rather than in clean skill spans. The ESCO mining component partly compensates for this by retrieving nearby taxonomy labels from text chunks, but incorrect mined labels can also inject weak evidence for generic skills.

Second, related-domain false positives remain a central risk. Candidates from compliance, operations, engineering support, or analytics roles can share many skills with the target occupation while still differing in day-to-day function. This motivates treating taxonomy features as calibration evidence: relatedness provides signal, but final relevance still requires the correct occupational function.

Third, seniority mismatch is not fully solved by embedding similarity. A senior managerial CV may mention the right domain and tools but be less appropriate for an individual contributor or support role, and the reverse can also occur. The listwise LLM prompt was therefore asked to consider role level and evidence quality, but the final score remains anchored to the retriever to avoid overreacting to one generative judgment.

Fourth, Spanish and cross-lingual matching introduce phrasing mismatches. The shared ESCO URI space and multilingual BGE-M3 embeddings reduce this problem, but Spanish CVs often contain different occupational phrasing, inflected forms, or locally conventional role descriptions. This makes exact skill spans and role labels less reliable than taxonomy-normalized evidence.

6. Demographic Masking and Ranking Stability

The submitted system uses demographic masking as an explicit procedural safeguard. Basic masking is applied to all scoring text to reduce direct use of contact cues and simple demographic markers, including emails, URLs, phone-like numbers, pronouns, and common honorifics or titles. The same masked text is used by the lexical retriever, embedding rankers, skill extraction/mining components, graph calibration, and LLM reranker, so the intervention is applied consistently in both the non-LLM and LLM variants. The LLM prompt also emphasizes job-relevant evidence such as skills, experience, certifications, and role requirements.

The available evaluation signal is the official controlled-bias RBO score, computed over the `en-es` and `es-es` directions. Table 5 reports these organizer-provided values. In the table, `Avg.` RBO averages the two annotated directions. We interpret the relatively high RBO values as evidence that the final ranking is relatively stable under the organizer's demographic comparison. This is a stability signal, not a proof of fairness or a disparate-impact evaluation.

Masking can only reduce direct demographic leakage. Indirect proxies can remain in work history, education, geography, language, career interruptions, or occupational trajectories, and masking can also remove relevant context in edge cases. Stronger fairness evaluation would require group-conditioned effectiveness metrics, counterfactual CV rewriting, and multi-objective tuning that balances MAP with

Table 5
Controlled-bias RBO scores (%) for the submitted variants.

System	RBO(en-es, %)	RBO(es-es, %)	Avg. RBO (%)
Retrieval + graph, no LLM	86.29	90.82	88.56
Final bounded LLM system	86.67	91.14	88.91

ranking stability. The submitted system should therefore be interpreted as using demographic masking and stability evaluation, not as solving fairness more broadly.

7. Limitations

The main experimental limitation is the small development set. The gap between development MAP and hidden-test MAP suggests distribution shift and possible overfitting to development queries, so development scores should be interpreted as directional ablation evidence rather than as hidden-test estimates. Local development sweeps showed that neighboring retrieval and fusion settings produced similar development MAP, but because these variants were not submitted, we do not claim hidden-test stability for them.

The hidden test labels prevent detailed test error analysis. The qualitative examples and module diagnostics are therefore based on development qrels and submitted-run logs, which limits how precisely we can attribute the remaining test errors to individual components.

The method also depends on ESCO mapping quality. Missing skill spans, noisy ESCO mining, or overly generic graph nodes can still produce false positives, and SkillMatch MPNet is most naturally aligned with English job-skill matching. Spanish and cross-lingual behavior therefore rely more heavily on BGE-M3, ESCO URI normalization, and LLM reranking.

Finally, LLM reranking adds cost, latency, and formatting risk. To keep the listwise prompt bounded, each CV is truncated to 1,800 characters with a head/tail split. This preserves typical summary and recent-history evidence, but it can miss relevant skills or certifications that appear only in the omitted middle of a long CV. Malformed or incomplete listwise outputs can also reduce the value of top-k refinement even when a complete retrieval ranking remains available.

8. Discussion and Conclusion

The official test results support a retrieval-first interpretation of TalentCLEF Task A. On the official test set, the retrieval-plus-graph system already reaches 57.62% MAP on the monolingual average and 48.64% MAP on `en-es`. Adding the bounded LLM reranker improves the monolingual average by 2.27 MAP percentage points and `en-es` by 4.29 points; on development, it adds 1.63 points over the graph-only system. These official deltas, together with the development diagnostics, are consistent with treating the LLM as an optional top-k refinement layer over a complete retrieved candidate pool, not as the primary ranking mechanism.

The ESCO layer shows a complementary diagnostic pattern. Normalized occupational evidence can recover related skills when exact labels differ, and IDF plus hub-node penalties reduce the influence of generic concepts. On development, soft graph calibration improves average MAP from 81.61% to 82.93%, mainly through the Spanish ranking, while the stricter role-gate variant drops to 81.33%. We therefore used ESCO evidence as calibration over retrieved candidates, not as an exclusion rule, in the submitted system.

Taken together, the findings support a practical multilingual HR ranking design: use robust retrieval for coverage, taxonomy calibration for occupational evidence, and LLM reranking only where it adds clear ordering value, most visibly in the cross-lingual setting.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI Codex as a writing and source-code inspection assistant in order to: reformulate, reorganize, and polish parts of the paper; inspect repository code; and prepare LaTeX edits based on feedback and instructions from the authors. The submitted system, experiments, results, and methodological choices were defined and validated by the authors. After using this tool, the authors reviewed and edited all generated text, verified the reported results against the submitted systems and evaluation outputs, and take full responsibility for the content of the paper.

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] 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] N. Otani, N. Bhutani, E. Hruschka, Natural language processing for human resources: A survey, in: W. Chen, Y. Yang, M. Kachuee, X.-Y. Fu (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 583–597. URL: https://aclanthology.org/2025.naacl-industry.47/. doi:10.18653/v1/2025.naacl-industry.47.

[5] E. Senger, M. Zhang, R. van der Goot, B. Plank, Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings, in: E. Hruschka, T. Lake, N. Otani, T. Mitchell (Eds.), Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1–15. URL: https://aclanthology.org/2024.nlp4hr-1.1/. doi:10.18653/v1/2024.nlp4hr-1.1.

[6] M. Zhang, K. Jensen, S. Sonniks, B. Plank, SkillSpan: Hard and soft skill extraction from English job postings, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4962–4984. URL: https://aclanthology.org/2022.naacl-main.366/. doi:10.18653/v1/2022.naacl-main.366.

[7] M. Zhang, R. van der Goot, B. Plank, ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11871–11890. URL: https://aclanthology.org/2023.acl-long.662/. doi:10.18653/v1/2023.acl-long.662.

[8] M. Zhang, R. van der Goot, B. Plank, Entity linking in the job market domain, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 410–419. URL: https://aclanthology.org/2024.findings-eacl.28/. doi:10.18653/v1/2024.findings-eacl.28.

[9] M. Zhang, R. van der Goot, M.-Y. Kan, B. Plank, NNOSE: Nearest neighbor occupational skill extraction, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European

Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 589–608. URL: https://aclanthology.org/2024.eacl-long.35/. doi:10.18653/v1/2024.eacl-long.35.

[10] K. Nguyen, M. Zhang, S. Montariol, A. Bosselut, Rethinking skill extraction in the job market domain using large language models, in: E. Hruschka, T. Lake, N. Otani, T. Mitchell (Eds.), Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 27–42. URL: https://aclanthology.org/2024.nlp4hr-1.3/. doi:10.18653/v1/2024.nlp4hr-1.3.

[11] A. Magron, A. Dai, M. Zhang, S. Montariol, A. Bosselut, JobSkape: A framework for generating synthetic job postings to enhance skill matching, in: E. Hruschka, T. Lake, N. Otani, T. Mitchell (Eds.), Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 43–58. URL: https://aclanthology.org/2024.nlp4hr-1.4/. doi:10.18653/v1/2024.nlp4hr-1.4.

[12] B. Clavié, G. Soulié, Large language models as batteries-included zero-shot esco skills matchers, 2023.

[13] H. Kavas, M. Serra-Vidal, L. Wanner, Multilingual skill extraction for job vacancy–job seeker matching in knowledge graphs, in: G. A. Gesese, H. Sack, H. Paulheim, A. Merono-Penuela, L. Chen (Eds.), Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK), International Committee on Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 146–155. URL: https://aclanthology.org/2025.genaik-1.15/.

[14] N. Li, B. Kang, T. De Bie, Building data-driven occupation taxonomies: A bottom-up multi-stage approach via semantic clustering and multi-agent collaboration, in: S. Potdar, L. Rojas-Barahona, S. Montella (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Suzhou (China), 2025, pp. 1596–1614. URL: https://aclanthology.org/2025.emnlp-industry.113/. doi:10.18653/v1/2025.emnlp-industry.113.

[15] J. Li, B. Xu, Z. Chen, C. Xu, M. Chen, S. Liu, Y. Zhou, Z. Wen, Enhancing talent search ranking with role-aware expert mixtures and LLM-based fine-grained job descriptions, in: S. Potdar, L. Rojas-Barahona, S. Montella (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Suzhou (China), 2025, pp. 185–194. URL: https://aclanthology.org/2025.emnlp-industry.13/. doi:10.18653/v1/2025.emnlp-industry.13.

[16] X. Yu, R. Xu, C. Xue, J. Zhang, X. Ma, Z. Yu, ConFit v2: Improving resume-job matching using hypothetical resume embedding and runner-up hard-negative mining, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 12775–12790. URL: https://aclanthology.org/2025.findings-acl.661/. doi:10.18653/v1/2025.findings-acl.661.

[17] H. Kavas, M. Serra-Vidal, L. Wanner, Using large language models and recruiter expertise for optimized multilingual job offer - applicant cv matching, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, 2024. URL: https://doi.org/10.24963/ijcai.2024/1011. doi:10.24963/ijcai.2024/1011.

[18] M. Zhang, R. van der Goot, Nlpnorth @ talentclef 2025: Comparing discriminative, contrastive, and prompt-based methods for job title and skill matching, 2025. URL: https://arxiv.org/abs/2506.19058. arXiv:2506.19058.

[19] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is ChatGPT good at search? investigating large language models as re-ranking agents, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 14918–14937. URL: https://aclanthology.org/2023.emnlp-main.923/. doi:10.18653/v1/2023.emnlp-main.923.

[20] X. Ma, X. Zhang, R. Pradeep, J. Lin, Zero-shot listwise document reranking with a large language model, 2023. URL: https://arxiv.org/abs/2305.02156. arXiv:2305.02156.

[21] M. Adeyemi, A. Oladipo, R. Pradeep, J. Lin, Zero-shot cross-lingual reranking with large language

models for low-resource languages, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 650–656. URL: https://aclanthology.org/2024.acl-short.59/. doi:10.18653/v1/2024.acl-short.59.

[22] H. Iso, P. Pezeshkpour, N. Bhutani, E. Hruschka, Evaluating bias in LLMs for job-resume matching: Gender, race, and education, in: W. Chen, Y. Yang, M. Kachuee, X.-Y. Fu (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 672–683. URL: https://aclanthology.org/2025.naacl-industry.55/. doi:10.18653/v1/2025.naacl-industry.55.

[23] N. Laosaengpha, T. Tativannarat, A. Rutherford, E. Chuangsuwanich, Mitigating language bias in cross-lingual job retrieval: A recruitment platform perspective, 2025. URL: https://arxiv.org/abs/2502.03220. arXiv:2502.03220.

[24] 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://zenodo.org/records/19652670. doi:10.5281/zenodo.19652670, dataset, published April 19, 2026. License: Creative Commons Attribution 4.0 International.

[25] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems 28 (2010) 20:1–20:38. doi:10.1145/1852102.1852106.

[26] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2024) 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[27] A. Bielinski, D. Brazier, From retrieval to ranking: A two-stage neural framework for automated skill extraction, in: Proceedings of the 5th Workshop on Recommender Systems for Human Resources, volume 4046 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4046/RecSysHR2025-paper_5.pdf.

[28] 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, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 758–759. URL: https://doi.org/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[29] European Commission, ESCO: European skills, competences, qualifications and occupations, 2026. URL: https://esco.ec.europa.eu/.

[30] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 855–864. URL: https://doi.org/10.1145/2939672.2939754. doi:10.1145/2939672.2939754.

[31] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.