Improving German VET Test Item Retrieval with ESCO-Grounded Knowledge Graph Re-Ranking

Alonso Palomino
mail@alonsopg.com

Abstract

Test item retrieval from item banks is an understudied component of exam assembly, particularly in vocational education, where targeted skills, domain-specific terminology, and hierarchical skill relations are not fully captured by distributional text embeddings alone. ESCO provides a structured taxonomy of skills, competences, and occupations. This paper proposes an ESCO-grounded knowledge graph re-ranker for German vocational education and training (VET) item retrieval. On a German benchmark, the strongest standalone ESCO-grounded re-ranker, using similarity-weighted direct ESCO links and 512-dimensional SVD representations, improves nDCG@50 by 29.2% over dense retrieval. A late-fusion hybrid with BGEGemma achieves the best overall result, 0.6700 nDCG@50 (+33.8%), and significantly outperforms both components. These results show that ESCO-grounded re-ranking adds complementary skill-aware signals for candidate item retrieval in exam assembly workflows.

1 Introduction

Retrieving test items from an item bank is an important upstream step in exam assembly, especially in vocational education where competency alignment is essential (van der Linden, 2005; Kurdi et al., 2020). The goal here is therefore to improve the candidate pool presented to exam designers, rather than to optimize final test construction under blueprint, fairness, or psychometric constraints. Existing lexical and neural ranking methods do not explicitly model relations among skills and competences. The European Skills, Competences, Qualifications and Occupations (ESCO) framework provides a multilingual representation of such relations and enables ontology-grounded similarity signals for German vocational education and training (VET) item retrieval.

Using the benchmark of Palomino et al. (2024),

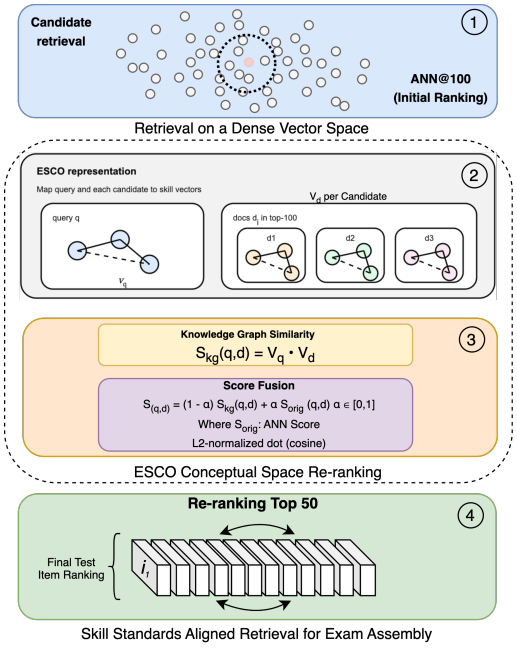

Figure 1: Base re-ranking pipeline: ANN retrieves the top-100 candidate test items; queries and candidates are mapped to ESCO-grounded skill representations to compute $s_{kg}(q, d) = v_q \cdot v_d$, which is fused with s_{orig} via α to re-rank and select the top-50.

the following research question is addressed: *To what extent can ESCO-based re-ranking improve the relevance of the top-k retrieved test items for German VET item retrieval, compared to lexical baselines and strong text-only re-rankers, including neural and LLM-based models?* (**RQ**). The best standalone variant reaches nDCG@50 = 0.6468, and a late-fusion hybrid with BGEGemma reaches 0.6700. These results indicate that ESCO-derived structured skill signals improve candidate item retrieval for exam assembly workflows and complement strong text-only re-ranking signals. This paper evaluates ESCO-based re-ranking for German VET item retrieval, a setting where ontology-grounded re-ranking has received little prior atten-

tion. The implementation is available online.[1]

2 Related Work

Test item retrieval from item banks is an important but understudied task in educational assessment (Lane et al., 2016; Kurdi et al., 2020). Prior work applies IR methods to distractor suggestion, item-bank management, and assessment pipelines (Ha and Yaneva, 2018; Liu et al., 2005; Zhou et al., 2019; Settles et al., 2020), while recent exam-assembly work ranks items using content, difficulty, and related facets (Palomino et al., 2024, 2025; Do et al., 2025). In parallel, ESCO and related taxonomies have been used in skill-centric NLP for multilingual pretraining, entity linking, skill extraction, framework mapping, and career-resource linking (Zhang et al., 2023, 2024, 2022; Magron et al., 2024; Kavas et al., 2024; Gnehm and Clematide, 2024; Senger et al., 2025). However, ESCO's graph structure has rarely been used as a direct re-ranking signal for German VET item retrieval.

3 Task and dataset

To answer **RQ**, the task is candidate item retrieval: given a query representing the skill set associated with a question in a VET item bank, the goal is to retrieve the most relevant test items from the existing item bank. For example, a paraphrased query may target calculating material costs and preparing an invoice; a relevant item would present quantities, prices, and customer-order information, while a non-relevant item might test customer-service etiquette or workplace safety vocabulary. ESCO helps by linking the query and relevant item to shared or nearby skills such as *calculate costs*, *prepare invoices*, and *use spreadsheet software*, while the non-relevant item links to a different skill neighborhood. Because retrieval is the first step in exam construction workflows, this section evaluates whether ESCO-based re-ranking improves top-k retrieval relevance over lexical, neural, and LLM-based baselines.

Testbed. Experiments use the non-proprietary fold of the benchmark of Palomino et al. (2024), built from a real German VET item bank used in exam assembly workflows (2,812 test items, 15 German VET queries, and relevance judgments). The benchmark is used to evaluate the proposed re-ranking methods, alternative lexical, neural, and LLM-based baselines, and the late-fusion hybrid extension.

ESCO Taxonomy. ESCO[2] is an EU multilingual taxonomy of skills, competences, qualifications, and occupations. Its hierarchical and associative links provide structured signals for knowledge-informed German VET test item retrieval.

ESCO Knowledge Graph Construction. A multi-relational knowledge graph (KG) is constructed from ESCO (version 23.1), yielding 13,939 entities and 12,530 triples across four relation types: broader, narrower, optional, and reverse optional. The broader/narrower relations encode the two directions of adjacent ESCO hierarchy links, while optional denotes an ESCO skill-skill association and reverse optional is the added reverse edge that lets graph methods traverse the association both ways. Figure 2 illustrates how a paraphrased VET query and positive/negative candidate items are linked to ESCO skills, and how these links help distinguish candidates through shared skills, different skill neighborhoods, and typed ESCO triples. Over the full ESCO skill inventory used as nodes, the resulting relation subset is sparse and long-tailed: 60.3% of nodes are isolated, 90.9% have degree at most 2, and only 17.8% fall in the largest connected component, limiting the local relational context available to graph-embedding models (Appendix A.5); construction details are reported in Appendix A.4.

After dense retrieval produces a top-100 candidate set, Figure 2 shows the re-ranking step: query and candidate texts are mapped to ESCO skills, shared cost/invoice skills provide higher KG support for the relevant item, and a service/safety item remains in a different skill neighborhood with lower KG support.

Knowledge Graph Representation Learning. Skill-aware representations are derived using ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), and SVD-based methods. The strongest setting uses similarity-weighted direct ESCO links without graph expansion and 512 latent dimensions; training and representation details are reported in Appendix A.4.

[1] https://github.com/alonsopg/ESCO-KGE-reranking

[2] https://esco.ec.europa.eu/en

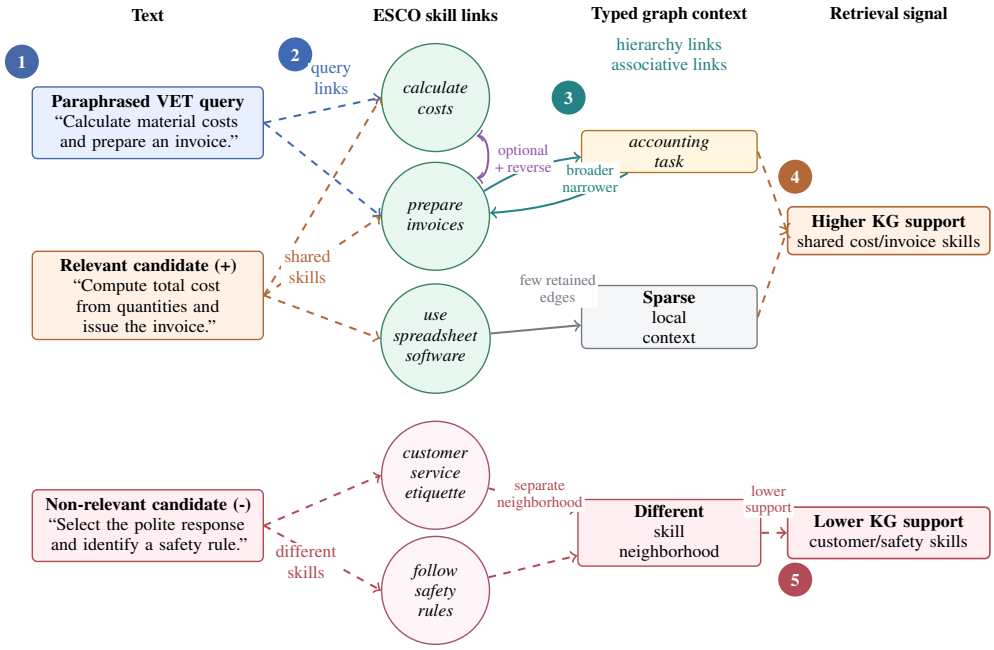

Figure 2: Illustrative ESCO-assisted retrieval graph with positive and negative candidates. After dense retrieval returns an initial candidate set, query and candidate texts are mapped to ESCO skill links; typed ESCO triples then provide graph context through hierarchy edges (broader/narrower) and associative skill-skill edges (optional, plus an added reverse optional edge for two-way traversal). A relevant item shares the query's cost/invoice skills and receives higher KG support, while a non-relevant item links to a different customer-service/safety neighborhood and receives lower support. The query, items, and labels are paraphrased for exposition.

4 Knowledge Graph Based Re-Ranking

To align retrieval with German vocational education and training (VET), the re-ranking approach combines dense retrieval with similarity signals derived from ESCO-based representations. For each query q, a German-language dense retriever[3] implemented with `sentence-transformers` (Reimers and Gurevych, 2020) and `FAISS` (Douze et al., 2024) retrieves a candidate set $\mathcal{C}_q$ of size $K_{cand} = 100$, yielding tuples $(q, d, s_{orig}(q, d))$, where d is a test item and s_{orig} is the original retriever score.

ESCO skill linking. To obtain ESCO-aligned representations, both queries and items are linked to ESCO skills using embedding-based nearest neighbor retrieval over ESCO German skill labels. All ESCO skills are indexed by their `preferredLabel` strings and encoded with the same SentenceTransformer encoder, building a FAISS inner-product index over L2-normalized label embeddings. Given a query text or an item text (stem plus answer options), the top-k nearest ESCO skills are retrieved and only those above a minimum similarity threshold are retained; each retained link is stored as an ESCO `conceptUri` with weight given by the cosine similarity. The linked skill nodes are then mapped to their precomputed ESCO knowledge-graph node embeddings (e.g., SVD/ComplEx/RotatE). In all experiments, queries (top-20) and items (top-15) are linked to ESCO skills with cosine ≥ 0.45, and v_q and v_d are computed as similarity-weighted means of the linked skill embeddings; if no skill is linked, the zero vector is used. This fixed threshold was chosen from preliminary ESCO nearest-neighbor inspection to prune low-similarity labels while preserving coverage, and was not tuned on retrieval relevance labels (Appendix A.2).

The strongest SVD setting uses direct similarity-weighted ESCO links without graph expansion, projected into a 512-dimensional latent space. After row-wise L2 normalization, knowledge graph similarity $s_{kg}(q, d)$ is computed by dot product, and the final score is obtained by convex interpolation:

$$s(q, d) = (1 - \alpha)\, s_{kg}(q, d) + \alpha\, s_{orig}(q, d), \quad (1)$$

where $\alpha \in [0, 1]$ weights s_{kg} vs. s_{orig}. For the strongest standalone SVD setting, a dedicated

[3] https://huggingface.co/deutsche-telekom/gbert-large-paraphrase-cosine

sweep over $\alpha \in [0.1, 0.9]$ found the best performance in the range $[0.25, 0.35]$, and $\alpha = 0.30$ is therefore reported for this variant. The graph-augmented SVD and KGE-based re-rankers retain $\alpha = 0.5$ as a simple reference setting. For the hybrid system, the best standalone ESCO-grounded re-ranker and BGEGemma are combined over the same dense candidate set after query-wise min-max normalization:

$$s_{hyb}(q,d) = \beta\, \tilde{s}_{svd}(q,d) + (1-\beta)\, \tilde{s}_{bge}(q,d), \quad (2)$$

where $\tilde{s}_{svd}$ and $\tilde{s}_{bge}$ denote normalized scores and $\beta \in [0, 1]$ controls the contribution of the ESCO-grounded SVD component. The strongest hybrid setting uses $\beta = 0.55$. Because no separate development split was available for this small benchmark, both α and β were selected by lightweight sweeps on the same 15-query testbed and are therefore reported as benchmark-specific tuned settings (Appendix A.1).

5 Experiments

Thirteen retrieval methods are evaluated, covering ESCO-based, hybrid KG+LLM, lexical, neural, and LLM-based re-ranking approaches. All runs are implemented and evaluated in `PyTerrier` (Macdonald and Tonellotto, 2020). For re-ranking, a top-$K_{cand} = 100$ candidate set is retrieved and the top-$K_{final} = 50$ is evaluated after re-ranking; the ANN baseline selects the top-$K_{final} = 50$ items directly from the dense retriever output. Here, *document* denotes a candidate test item.

Re-ranker inference. BGEGemma and the cross-encoder are used zero-shot, without VET-specific fine-tuning. BGEGemma uses `BAAI/bge-reranker-v2-gemma`, the Gemma-2B-based multilingual BGE v2 LLM reranker. Its template serializes pairs as `A: <query>` and `B: <document>`, then asks whether `B` answers `A` with `Yes` or `No`; the score is the next-token `Yes` logit, so no generation or decoding parameters (sampling, beams, top-p, temperature) are used. The cross-encoder uses `cross-encoder/msmarco-MiniLM-L6-en-de-v1`, an EN-DE MiniLM cross-encoder trained on MS MARCO passage ranking; query-document pairs are tokenized jointly with maximum length 512 and ranked by the scalar sequence-classification score.

Compared Methods:

1. **Hybrid (ESCO-grounded SVD direct-link, 512, $\alpha = 0.30$ + BGEGemma):** A late-fusion hybrid combines the best standalone ESCO-grounded re-ranker with BGEGemma over the shared dense top-100 candidate set using query-wise normalized scores, with $\beta = 0.55$ for the ESCO-grounded component.
2. **ESCO-grounded SVD re-ranker (direct-link, 512):** Similarity-weighted document-skill and query-skill matrices are built from direct ESCO links without graph expansion; truncated SVD with 512 components is fitted on the document-skill matrix, and latent similarity is combined with the original dense score using $\alpha = 0.30$.
3. **BGEGemma:** This LLM-based re-ranker (Li et al., 2023; Chen et al., 2024)[4] scores each dense query-document pair with the fixed BGEGemma relevance-scoring setup above.
4. **KG re-ranker (SVD graph-augmented, $\alpha = 0.5$):** Sparse document-skill and query-skill matrices are built from KG-augmented ESCO links; truncated SVD is fitted on the document-skill matrix, and latent similarity is combined with the original dense score.
5. **KG re-ranker (ComplEx):** ComplEx skill representations are aggregated into query and document vectors, whose similarity is combined with the original dense score.
6. **KG re-ranker (RotatE):** RotatE skill representations are aggregated into query and document vectors, whose similarity is combined with the original dense score.
7. **ANN@100→@50:** The top-50 items are selected directly from the dense retriever's top-100 output without re-ranking.
8. **Cross-encoder:** The EN-DE MiniLM cross-encoder[5] scores each dense query-document pair with its sequence-classification relevance score.
9. **Multilingual learned sparse retriever (MLSR):** Because standard sparse baselines such as SPLADE are primarily English/MS MARCO-trained, MLSR serves as the German-compatible sparse neural baseline (Geng et al., 2024)[6] and re-scores dense candidates with sparse dot-product similarity over lexical-semantic representations.

[4] `https://huggingface.co/BAAI/bge-reranker-v2-gemma`
[5] `https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1`
[6] Hugging Face model card.

#	Method	nDCG	Δ	%	MRR	P	R	F1	MAP
1	**Hybrid (ESCO-SVD direct-link + BGEGemma)**	**0.6700**	**+0.1694**	**+33.8**	**0.9407**	**0.5760**	**0.4925**	**0.5310**	**0.4349**
2	ESCO-SVD re-ranker (direct-link)	0.6468	+0.1462	+29.2	0.9042	0.5613	0.4814	0.5183	0.4105
3	BGEGemma	0.6008	+0.1002	+20.0	0.9111	0.5187	0.4521	0.4831	0.3667
4	KG re-ranker (SVD graph-aug.)	0.5986	+0.0980	+19.6	0.9200	0.5267	0.4590	0.4905	0.3504
5	KG re-ranker (ComplEx, $\alpha = 0.5$)	0.5516	+0.0510	+10.2	0.8222	0.4787	0.4247	0.4501	0.3083
6	KG re-ranker (RotatE, $\alpha = 0.5$)	0.5497	+0.0491	+9.8	0.8000	0.4773	0.4239	0.4490	0.3058
7	ANN@100→@50	0.5006	+0.0000	+0.0	0.7944	0.4627	0.4113	0.4355	0.2734
8	Cross-encoder	0.4650	-0.0356	-7.1	0.6811	0.4507	0.3973	0.4223	0.2367
9	MLSR	0.4470	-0.0536	-10.7	0.6861	0.4360	0.3797	0.4059	0.2097
10	DPH	0.3625	-0.1381	-27.6	0.7092	0.3320	0.2148	0.2608	0.1301
11	TF-IDF	0.3570	-0.1436	-28.7	0.6978	0.3293	0.2121	0.2580	0.1240
12	PL2	0.3538	-0.1468	-29.3	0.7028	0.3273	0.2100	0.2558	0.1232
13	BM25	0.3179	-0.1827	-36.5	0.6656	0.3107	0.1912	0.2367	0.1066

Table 1: Retrieval results at cutoff 50. Δ and % denote absolute and relative changes in nDCG@50 compared to the dense baseline (ANN@100→@50).

10. **DPH:** A Divergence From Randomness model that re-ranks candidates by deviation from a random term distribution.
11. **TF-IDF:** Re-ranks dense candidates by term-overlap similarity weighted by corpus specificity.
12. **PL2:** A DFR model with Poisson approximation and Laplace smoothing that re-ranks candidates by lexical relevance.
13. **BM25:** Re-ranks the dense top-100 candidates with a standard term-matching score.

5.1 Results

Table 1 reports retrieval performance at cutoff 50. The strongest overall method is the Hybrid (ESCO-grounded SVD direct-link, 512, $\alpha = 0.30$ + BGEGemma), which reaches nDCG@50 = 0.6700 (Δ=+0.1694, +33.8%) and yields the largest MAP@50 gain (0.4349 vs. 0.2734). Among standalone methods, ESCO-grounded SVD re-ranker (direct-link, 512, $\alpha = 0.30$) is best at 0.6468 (Δ=+0.1462, +29.2%), followed by BGEGemma at 0.6008 and the graph-augmented SVD setting at 0.5986; ComplEx and RotatE reach 0.5516 and 0.5497. Among the remaining text-only neural baselines, the cross-encoder reaches 0.4650 and MLSR 0.4470, while DPH remains the strongest lexical re-ranker at 0.3625. The cross-encoder's weak result likely reflects domain mismatch, as the EN-DE MiniLM model is general-domain and used zero-shot, whereas BGEGemma provides a substantially stronger text-only neural baseline in the same setting. The stronger direct-link@512 standalone result suggests that higher-rank latent compression of similarity-weighted skill assignments is more beneficial here than one-hop graph expansion, showing that graph expansion is not necessary for the best standalone ESCO-grounded result and reinforcing the view that SVD benefits from modeling global document-skill co-occurrence over a sparse, long-tail ESCO graph (Appendix A.5). The hybrid further indicates that ESCO-grounded and LLM-based re-ranking signals are complementary rather than redundant: it significantly outperforms both the best standalone ESCO-grounded re-ranker and BGEGemma on per-query nDCG@50 (Appendix A.1). A focused ablation over SVD rank and weighting is reported in Appendix A.3; this ablation is conducted under a fixed-α setting, while the standalone headline result additionally benefits from the later $\alpha = 0.30$ tuning. Under the shared ESCO linking configuration, linking covers 100.0% of queries and 99.6% of documents, with no zero-link queries and only 12 zero-link documents, suggesting that the ESCO-grounded gains are unlikely to reflect widespread linking failures; full diagnostics are reported in Appendix A.2.

6 Conclusions

To address **RQ**, lexical, neural, ESCO-based, and KG+LLM hybrid methods were compared for German VET item retrieval. The ESCO-based variants improve over dense retrieval and lexical re-rankers without manual skill annotation; the strongest standalone SVD variant raises nDCG@50 from 0.5006 to 0.6468. Late fusion with BGEGemma further reaches 0.6700 and significantly outperforms both standalone signals, indicating that ESCO-grounded and LLM-based re-ranking are complementary. Overall, structured skill-aware signals improve candidate retrieval for exam assembly.

Limitations and Ethics Statement

This work addresses candidate item retrieval as an upstream support step in exam assembly. The system improves access to existing items, but it does not optimize final test construction under blueprint, difficulty, exposure, fairness, or psychometric constraints, and it does not generate new test items. The results should therefore be interpreted as evidence that ESCO-grounded signals can improve the candidate pool presented to exam designers, not as evidence that the resulting ranked list is itself a complete exam or a substitute for expert assembly decisions. Because the benchmark contains 15 queries, the results should be read as evidence on this testbed rather than as a final estimate across VET retrieval settings.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, Jonghwi Kim, and Gary Lee. 2025. Multi-facet blending for faceted query-by-example retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28577–28590, Vienna, Austria. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Zhichao Geng, Dongyu Ru, and Yang Yang. 2024. Towards competitive search relevance for inference-free learned sparse retrievers. *arXiv preprint arXiv:2411.04403*.

Ann-Sophie Gnehm and Simon Clematide. 2024. Mapping work task descriptions from German job ads on the O*NET work activities ontology. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11049–11059, Torino, Italia. ELRA and ICCL.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398, New Orleans, Louisiana. Association for Computational Linguistics.

Hamit Kavas, Marc Serra-Vidal, and Leo Wanner. 2024. Enhancing job posting classification with multilingual embeddings and large language models. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 440–450, Pisa, Italy. CEUR Workshop Proceedings.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Suzanne Lane, Mark R. Raymond, and Thomas M. Haladyna, editors. 2016. *Handbook of Test Development*, 2 edition. Routledge, New York, NY.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*.

Wim J. van der Linden. 2005. *Linear Models for Optimal Test Design*. Springer, New York, NY.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. JobSkape: A framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 43–58, St. Julian's, Malta. Association for Computational Linguistics.

Alonso Palomino, Andreas Fischer, David Buschhüter, Roland Roller, Niels Pinkwart, and Benjamin Paassen. 2025. Mitigating bias in item retrieval for enhancing exam assembly in vocational education services. In *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)*, pages 183–193, Albuquerque, New Mexico. Association for Computational Linguistics.

Alonso Palomino, Andreas Fischer, Jakub Kuzilek, Jarek Nitsch, Niels Pinkwart, and Benjamin Paassen. 2024. EdTec-QBuilder: A semantic retrieval tool for assembling vocational training exams in German language.

In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 26–35, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elena Senger, Yuri Campbell, Rob van der Goot, and Barbara Plank. 2025. KARRIEREWEGE: A large scale career path prediction dataset. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 533–545, Abu Dhabi, UAE. Association for Computational Linguistics.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2023. ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11871–11890, Toronto, Canada. Association for Computational Linguistics.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Entity linking in the job market domain. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 410–419, St. Julian's, Malta. Association for Computational Linguistics.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. 2022. SkillSpan: Hard and soft skill extraction from English job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984, Seattle, United States. Association for Computational Linguistics.

Wei Zhou, Renfen Hu, Feipeng Sun, and Ronghuai Huang. 2019. An intelligent testing strategy for vocabulary assessment of Chinese second language learners. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 21–29, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Significance tests

Per-query nDCG@50 differences over 15 queries are tested using paired randomization (sign-flip), a paired t-test, and the Wilcoxon signed-rank test. Because no separate development split was available for this small benchmark, α and β were selected by lightweight sweeps on the same 15-query testbed and are therefore reported as benchmark-specific tuned settings. Table 2 reports p values and the mean difference Δ (system minus baseline) for the strongest standalone ESCO-grounded re-ranker and for the hybrid ESCO-grounded plus LLM re-ranker.

A.2 Linking diagnostics

Unit	Top-k	Min sim.	Cov. (%)	Zero	Full (%)	Avg.	Med.	p10	p90
Query	20	0.45	100.0	0	100.0	20.0	0.713	0.642	0.815
Document	15	0.45	99.6	12	94.2	14.6	0.608	0.517	0.713

Table 3: Diagnostics for the ESCO skill-linking stage under the shared ESCO linking configuration. Cov. denotes coverage, Zero the number of zero-link cases, Full the percentage retaining the full top-k, Avg. the average retained links, and Med. the median similarity.

To further examine the behavior of the linking stage, Table 3 reports coverage, zero-link frequency, retained-link counts, and similarity-score distributions under the shared ESCO linking configuration. ESCO linking covers 100.0% of queries and 99.6% of documents, with no zero-link queries and only 12 zero-link documents in the full collection. The current threshold is also permissive rather than brittle: 100.0% of queries retain the full top-20 skills, and 94.2% of documents retain the full top-15, with documents keeping 14.6 links on average. Similarity scores remain consistently above the threshold, with median cosine similarity of 0.713 for queries and 0.608 for documents. Although gold skill annotations are not available to compute linking precision and recall directly, these diagnostics provide complementary evidence that the ESCO linking stage is high-coverage and that the retrieval gains are not driven by frequent linking failures.

Comparison	Δ	Rand. p	t p	W p
ESCO-grounded SVD direct-link (512, $\alpha = 0.30$) vs ANN@100→@50	+0.1462	0.00035	0.00007	0.00031
ESCO-grounded SVD direct-link (512, $\alpha = 0.30$) vs Cross-encoder	+0.1817	0.00025	0.00015	0.00031
ESCO-grounded SVD direct-link (512, $\alpha = 0.30$) vs BGEGemma	+0.0469	0.07300	0.07185	0.10942
Hybrid (ESCO-grounded SVD + BGEGemma) vs ESCO-grounded SVD direct-link (512, $\alpha = 0.30$)	+0.0232	0.01860	0.02659	0.02194
Hybrid (ESCO-grounded SVD + BGEGemma) vs BGEGemma	+0.0701	0.00275	0.00309	0.00429
Hybrid (ESCO-grounded SVD + BGEGemma) vs ANN@100→@50	+0.1694	0.00030	0.00002	0.00031
Hybrid (ESCO-grounded SVD + BGEGemma) vs Cross-encoder	+0.2048	0.00005	0.00001	0.00006

Table 2: Paired significance tests on per-query nDCG@50 over 15 queries. Δ is the mean nDCG@50 difference (system minus baseline). Rand. p is a paired randomization (sign-flip) test; t p is a paired t-test; W p is the Wilcoxon signed-rank test. The standalone SVD rows correspond to the tuned $\alpha = 0.30$ setting reported in the main paper.

A.3 SVD rank and weighting ablation

Variant	Weighting	Graph	Rank	nDCG@50	Δ vs ANN
Graph-augmented ref.	Similarity	1-hop	256	0.5986	+0.0980
Best sim+graph	Similarity	1-hop	384	0.6397	+0.1391
Best sim-only	Similarity	None	512	0.6401	+0.1395
Best binary+graph	Binary	1-hop	384	0.6147	+0.1141
Best binary-only	Binary	None	512	0.6368	+0.1362

Table 4: Focused ablation over SVD rank and skill-assignment weighting in the dense@100→rerank→@50 setup under a fixed interpolation weight $\alpha = 0.5$. The former graph-augmented SVD configuration is included for reference; the remaining rows report the best variant within each weighting family. All best-per-family variants exceed the ComplEx (0.5516) and RotatE (0.5497) reference results.

To further probe why SVD outperforms ComplEx and RotatE, Table 4 reports a focused ablation over latent rank and skill-assignment weighting under a fixed $\alpha = 0.5$. First, the former graph-augmented configuration (similarity-weighted links with 1-hop expansion at rank 256) reaches 0.5986 nDCG@50, confirming that the earlier reported system is reproducible within the same sweep. Second, the strongest standalone SVD setting under this fixed-α analysis is the similarity-weighted direct-link variant without graph expansion at rank 512, which reaches 0.6401; the best graph-augmented similarity-weighted variant is nearly as strong at 0.6397. Third, higher latent ranks consistently help, with the strongest results appearing at 384–512 dimensions, while even the best binary variants remain well above the ComplEx and RotatE references. A later interpolation sweep further improves this standalone sim-only variant to 0.6468 at $\alpha = 0.30$, which is the setting reported in the main paper. Taken together, these results suggest that SVD's advantage is robust across reasonable weighting choices and is driven more by latent compression of the skill-assignment structure than by local graph propagation alone; the main paper's best overall result then comes from late fusion of this strongest standalone ESCO-grounded signal with BGEGemma.

A.4 KG construction and representation details

A multi-relational knowledge graph is constructed from the ESCO taxonomy (version 23.1, released 2023-12-20). Each node represents a skill concept identified by a unique URI; human-readable labels (the *preferredLabel* property) are stored as metadata for indexing and are not part of the graph structure. Edges are derived from two ESCO resources: the *skills hierarchy*, which provides directed edges (*broader*, *narrower*) between adjacent taxonomy levels (Levels 0–3), and the *skill-skill relations* file, which provides non-taxonomic links represented as an original *optional* edge plus an added reverse edge labeled *optional_rev*. For downstream use, the graph is stored both as typed triples $(head, relation, tail)$ for embedding-based methods and as an undirected adjacency map for neighborhood expansion with exponential-decay weighting.

Skill-aware representations are derived using three ESCO-based methods. ComplEx and RotatE are trained on ESCO triples with `PyKEEN` (Ali et al., 2021) under a standard link-prediction objective, using 512 dimensions for ComplEx and 256 for RotatE; complex-valued embeddings are converted to real vectors by concatenating real and imaginary parts. For the SVD family, sparse document-skill and query-skill matrices are built from ESCO-linked skills, with both graph-augmented and direct-link variants. Truncated SVD is fitted on the document-skill matrix to obtain a shared latent skill space, and queries are projected with the

same transformation. Unlike ComplEx and RotatE, which learn directly from ESCO triples, SVD factorizes structured skill assignments; the strongest direct-link setting uses similarity-weighted ESCO links without graph expansion and 512 latent dimensions, explaining 77% of the variance. For ComplEx and RotatE, query and document vectors are computed as confidence-weighted averages of linked skill embeddings. KGE-based vectors use power normalization ($p = 0.5$) followed by L2 normalization, while SVD-based vectors use L2-normalized projections.

A.5 ESCO graph sparsity profile

Statistic	Value
Isolated nodes	8,406 (60.3%)
Leaf nodes (degree = 1)	2,971 (21.3%)
Nodes with degree ≤ 2	12,668 (90.9%)
Median degree	0
90th percentile degree	2
99th percentile degree	7
Maximum degree	25
Largest connected component	2,478 nodes (17.8%)

Table 5: Degree statistics for the 13,939-node, 12,530-triple ESCO graph used in the experiments.

Table 5 shows the sparse local graph profile: 60.3% of skills are isolated, 90.9% have degree at most 2, and only 17.8% fall in the largest component. Because edges are restricted to the four relation types used in the experiments, isolation means no retained task-specific ESCO edge. This explains why local KGEs are constrained by sparse neighborhoods, whereas SVD exploits global document-skill co-occurrence.