WorkRB: A Community-Driven Evaluation Framework for AI in the Work Domain

Matthias De Lange*
TechWolf
Ghent, Belgium

Warre Veys
TechWolf
Ghent, Belgium

Federico Retyk
Avature
Barcelona, Spain

Daniel Deniz
Avature
Barcelona, Spain

Warren Jouanneau
Malt
Paris, France

Mike Zhang
University of Copenhagen
Copenhagen, Denmark

Aleksander Bielinski
Edinburgh Napier University
Edinburgh, UK

Emma Jouffroy
Malt
Paris, France

Nicole Clobes
WAPES
Brussels, Belgium

Nina Baranowska
Leiden University
Leiden, Netherlands

David Graus
University of Amsterdam
Amsterdam, Netherlands

Marc Palyart
Malt
Paris, France

Rabih Zbib
Avature
Barcelona, Spain

Dimitra Gkatzia
Edinburgh Napier University
Edinburgh, UK

Thomas Demeester
Ghent University - imec
Ghent, Belgium

Tijl De Bie
AIDA-IDLab, Ghent University
Ghent, Belgium

Toine Bogers
IT University of Copenhagen
Copenhagen, Denmark

Jens-Joris Decorte
TechWolf
Ghent, Belgium

Jeroen Van Hautte
TechWolf
Ghent, Belgium

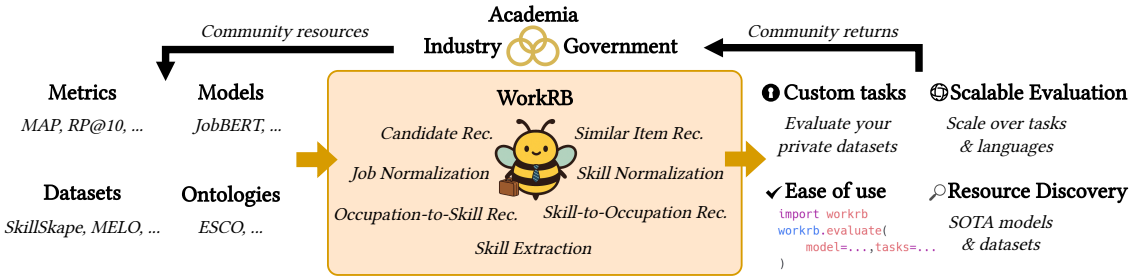

Figure 1: Overview of the community-driven WorkRB evaluation framework.

Abstract

Today's evolving labor markets rely increasingly on recommender systems for hiring, talent management, and workforce analytics, with natural language processing (NLP) capabilities at the core. Yet, research in this area remains highly fragmented. Studies employ divergent ontologies (ESCO, O*NET, national taxonomies), heterogeneous task formulations, and diverse model families, making cross-study comparison and reproducibility exceedingly difficult. General-purpose benchmarks lack coverage of work-specific tasks, and the inherent sensitivity of employment data further limits open

*Correspondence to workrb@techwolf.ai.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

RecSys '26, TBD

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2026/09
https://doi.org/XXXXXXX.XXXXXXX

evaluation. We present **WorkRB** (Work Research Benchmark), the first open-source, community-driven benchmark tailored to work-domain AI. WorkRB organizes 13 diverse tasks from 7 task groups as unified recommendation and NLP tasks, including job / skill recommendation, candidate recommendation, similar item recommendation, and skill extraction and normalization. WorkRB enables both monolingual and cross-lingual evaluation settings through dynamic loading of multilingual ontologies. Developed within a multi-stakeholder ecosystem of academia, industry, and public institutions, WorkRB has a modular design for seamless contributions and enables integration of proprietary tasks without disclosing sensitive data. WorkRB is available under the Apache 2.0 license at `https://github.com/techwolf-ai/WorkRB`.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

evaluation framework, recommendation benchmark, work domain, human resources, labor market intelligence, skill recommendation, job recommendation, multilingual evaluation, open-source

ACM Reference Format:

Matthias De Lange, Warre Veys, Federico Retyk, Daniel Deniz, Warren Jouanneau, Mike Zhang, Aleksander Bielinski, Emma Jouffroy, Nicole Clobes, Nina Baranowska, David Graus, Marc Palyart, Rabih Zbib, Dimitra Gkatzia, Thomas Demeester, Tijl De Bie, Toine Bogers, Jens-Joris Decorte, and Jeroen Van Hautte. 2026. WorkRB: A Community-Driven Evaluation Framework for AI in the Work Domain. In *Proceedings of Proceedings of the 20th ACM Conference on Recommender Systems (RecSys '26).* ACM, New York, NY, USA, 5 pages. `https://doi.org/XXXXXXX.XXXXXXX`

1 Introduction

Work-domain AI capabilities, including job–skill recommendation, skill extraction, job normalization, and related tasks, power hiring platforms, talent management systems, and public employment services worldwide. However, research in this area remains highly fragmented: practitioners and researchers employ a wide variety of task formulations, languages, model families, and ontologies such as ESCO [13], O*NET [17], SkillsFuture [21], and numerous national taxonomies. This fragmentation makes it exceedingly difficult to compare results across studies, reproduce findings, or build incrementally on prior work, slowing collective progress. Each task typically requires its own bespoke evaluation setup, from dataset preparation to metric selection, prohibiting scaling across multiple work-domain tasks. Moreover, work-domain data is inherently highly sensitive; career histories, compensation records, and employment data constitute personal data subject to strict privacy regulation, making it both undesirable and difficult to share datasets openly, which further limits reproducibility. Meanwhile, multilingual occupational ontologies such as ESCO contain rich hierarchical and cross-lingual structures, that remain largely untapped for evaluation purposes.

While common NLP benchmarks such as MTEB [15], BEIR [22], and SuperGLUE [23] focus on general-purpose capabilities, they are not aligned with recommendation tasks, evaluation strategies, and models specific to the work domain.

To this end, we introduce **WorkRB** (Work Research Benchmark), an open-source benchmark framework that unifies datasets, baseline models, and multilingual evaluation across work-related recommendation and NLP tasks. Our contributions are as follows:

(1) **Unified benchmark.** WorkRB scales evaluation across 13 recommendation and NLP tasks and 7 task groups, evaluated in a unified fashion as ranking problems [2]. Through dynamic ontology loading, WorkRB currently supports up to 28 languages, matching ESCO's full language coverage, enabling both monolingual and cross-lingual tasks.
(2) **Extensible toolkit.** WorkRB is pip-installable and features a modular design with extendable base classes for tasks and models, automatic checkpointing, and hierarchical metric aggregation. Open-source evaluation tasks can be complemented by proprietary, sensitive-data tasks for internal use.
(3) **Multi-stakeholder ecosystem.** As exemplified in Figure 1, WorkRB is a joint effort spanning academia, industry, and public institutions, bridging communities that each hold complementary pieces of the work-domain puzzle. Industrial and academic contributors provide real-world tasks, datasets, and domain-specialized models from published research; and government institutions supply extensive occupational ontologies (e.g., ESCO by the European Commission, O*NET by the US Department of Labor) that serve as the taxonomic backbone for many tasks. WorkRB provides a structured contribution model that incentivizes continued participation across all three communities.
(4) **Broader impact.** We discuss how open standardization supports legal compliance for sensitive work data, improves language representativeness, and enables standardized evaluation for public employment services.

2 The Work Research Benchmark

WorkRB[1] is designed to accommodate multiple ontologies, task types, and languages within a comparable evaluation framework. To avoid restricting evaluation to a single setup, WorkRB operates as a repository-style evaluation toolbox, providing a curated collection of datasets and baselines, with a focus on flexible configurations, rather than functioning as a constrained competition platform with a live leaderboard.

2.1 Evaluation Tasks & Ontologies

WorkRB comprises 13 recommendation and NLP tasks organized into 7 task groups, addressing core retrieval and ranking scenarios in the work domain. Following prior work [2], all of the contributed work-domain tasks are formulated as ranking problems. However, WorkRB remains flexible in architecture, providing support for classification tasks, and can also be extended into other task formats. Many tasks are grounded in an ontology, such as ESCO [13], for which WorkRB provides a flexible interface that automatically processes and caches the ontology for a given version and language, enabling efficient re-use across tasks. Table 1 provides a complete overview of all tasks with their label type, number of supported

[1]WorkRB is under continuous development by community contributions. The version described in this paper is based on *v0.5.1*.

languages, and dataset size in terms of their number of queries and targets for recommendation. We describe each task group below.

- **Occupation-to-Skill Recommendation (SRec)** ranks ESCO skills for a given occupation, where each occupation maps to multiple relevant skills [2, 13].
- **Skill-to-Occupation Recommendation (ORec)** ranks ESCO occupations for a given skill, where each skill maps to multiple relevant occupations [2, 13].
- **Similar Item Recommendation (SIRec)** ranks skills [6] or occupation titles [8, 24] by semantic relatedness, with single and multi-label targets respectively.
- **Candidate Recommendation (CRec)** ranks candidate profiles in cross-lingual settings given a freelancer project description or keyword search query, with multiple relevant candidates per query [11, 12].
- **Job Normalization (JNorm)** maps free-text job titles or national taxonomy entries to standardized ESCO occupation entries, including cross-lingual entity linking [5, 19].
- **Skill Normalization (SNorm)** maps skill surface forms or national taxonomy entries to canonical ESCO skill entries, including cross-lingual entity linking [2, 13, 19].
- **Skill Extraction (SExtr)** retrieves relevant skills from job descriptions, where queries are sentences ranked against a target skill taxonomy [4, 14].

2.2 Models & Baselines

The WorkRB toolkit provides inference implementations of diverse models and baselines for recommendation and NLP tasks, spanning both neural and traditional retrieval paradigms. On the semantic side, the benchmark provides a `BiEncoder` wrapper for any sentence-transformers model [18], along with domain-specialized models contributed from published research, including JobBERT-v1 to v3 [3, 5, 7] and CurriculumMatch [1], and token-level embedding models such as ConTeXTMatch [7]. For lexical baselines, WorkRB includes BM25 [20], TF-IDF, and Edit Distance as traditional retrieval reference points, complemented by a random ranking baseline that establishes a lower-bound reference. All models support adaptive target spaces, enabling consistent evaluation across diverse tasks. While the current release focuses on one-stage embedding-based models, the framework architecture is designed to support broader model families, including multi-stage and generative approaches. Table 2 reports mean average precision (MAP) scores across all task groups for a representative subset of baselines.

Table 1: Overview of WorkRB tasks. Dataset sizes are shown for the English (en) number of queries × targets. For non-English datasets, we separately report the maximum number of queries and targets for cross-lingual (x), Bulgarian (bg), and Swedish (sv). ESCO targets vary by language and version.

Dataset	Labels	Size (en)	Lang.
Occupation-to-Skill Rec.			
ESCO Occupation-to-Skill [2, 13]	multi	3,039 (en) × 13,939 (en)	28
Skill-to-Occupation Rec.			
ESCO Skill-to-Occupation [2, 13]	multi	13,492 (en) × 3,039 (en)	28
Similar Item Rec.			
Job Title Sim. [8, 24]	multi	105 (en) × 2,619 (en)	11
SkillMatch-1K [6]	single	900 (en) × 2,648 (en)	1
Candidate Rec.			
Query-Candidate [12]	multi	200 (en) × 4,019 (x)	5
Project-Candidate [12]	multi	200 (en) × 4,019 (x)	5
Job Normalization			
JobBERT [5]	single	15,463 (en) × 2,942 (en)	24
MELO (48 datasets) [19]	multi	4,438 (bg) × 150,140 (x)	21
Skill Normalization			
ESCO Alternatives [2, 13]	multi	72,008 (en) × 13,939 (en)	28
MELS (8 datasets) [19]	multi	4,381 (sv) × 100,273 (en)	5
Skill Extraction			
House [4]	multi	262 (en) × 13,891 (en)	28
Tech [4]	multi	338 (en) × 13,891 (en)	28
SkillSkape [14]	multi	1,191 (en) × 13,891 (en)	28

Table 2: MAP per task group. *Multilingual* models are evaluated on all languages with no language aggregation; *EN* models are evaluated on English monolingual subsets only. Best result for both scenarios in bold.

Model	SRec	ORec	SIRec	CRec	JNorm	SNorm	SExtr	All
#Multilingual datasets	*28*	*28*	*12*	*12*	*72*	*35*	*84*	*271*
Random Ranking	0.4	0.6	0.7	8.2	0.3	0.1	0.1	1.5
BM25 [20]	2.9	6.1	11.0	27.5	2.5	47.4	5.5	14.7
Qwen3-Embed (0.6B) [25]	6.0	14.7	26.3	**50.4**	12.5	**65.5**	**16.9**	27.5
JobBERT-v3 [3]	**9.3**	**20.5**	**30.6**	43.4	**25.6**	62.5	16.6	**29.8**
#EN-only datasets	*1*	*1*	*2*	*2*	*2*	*1*	*3*	*12*
ConTeXTMatch [7]	14.6	32.2	31.8	54.5	36.7	86.8	46.7	43.3
JobBERT-v2 [7]	15.4	33.2	**37.4**	**56.8**	**38.9**	83.2	27.5	41.8
CurriculumMatch [1]	**15.8**	**34.1**	34.4	52.1	36.7	**88.2**	**47.3**	**44.1**

2.3 Multilingual Support

Beyond multiple tasks, work-domain applications typically require support for multiple languages. Scaling evaluation over the quadratic task-language evaluation matrix is challenging as query and target spaces may be defined in different languages, ontologies carry distinct multilingual structures, and evaluation must account for both monolingual and cross-lingual settings. WorkRB addresses this by dynamically loading ontology structures per language and version, resolving query and target spaces independently. This allows users to define arbitrary monolingual setups (e.g., German queries and targets) or cross-lingual setups (e.g., French queries against English targets) within the same task. For example, the ESCO ontology [13] provides occupation and skill vocabularies in up to 28 languages, all of which WorkRB can load on demand for any ESCO-based tasks. Additionally, this design allows support for inherent cross-lingual tasks such as the candidate recommendation tasks operating across five languages simultaneously [11, 12], while MELO and MELS [19] provide cross-lingual normalization datasets that map national taxonomy entries to ESCO across 21 and 5 languages, respectively. Users can further define monolingual and cross-lingual metric aggregation strategies, as discussed in Section 3.

3 Design & Usage

Usage. WorkRB follows a simple three-step workflow: (1) initialize a model, (2) select tasks and languages, and (3) run evaluation.

Listing 1: WorkRB usage example for model evaluation.

```
from workrb import models, tasks, evaluate

model = models.BiEncoderModel("all-MiniLM-L6-v2")
tasks = [
 tasks.ESCOSkillNormRanking(split="val", languages=["en"]),
 tasks.MELORanking(split="val", languages=["de", "fr"])]
results = evaluate(model, tasks)
```

The framework is installable via "`pip install workrb`"; Listing 1 illustrates the core workflow.

Extensibility. A registry system ensures that all tasks and models are dynamically discoverable, where organizations can locally extend WorkRB with proprietary datasets and models, without the need to share sensitive data. Additionally, WorkRB is designed for sustained collaborative development, as detailed in Section 4.1. Extensibility is achieved through abstract base classes. Custom tasks inherit from `RankingTask`, implement `load_dataset()`, and register via `@register_task()`. Tasks map languages to dataset identifiers through `languages_to_dataset_ids()`, supporting multiple datasets per language for configurations with same-language datasets (e.g. across regions), and cross-lingual evaluation. Similarly, custom models inherit from `ModelInterface`, implement `compute_rankings()`, and register via `@register_model()`.

Checkpointing. WorkRB provides automatic checkpointing, saving results after each task–dataset completion. Checkpoints track `(task, dataset_id)` tuples, so re-running with the same output folder seamlessly resumes from where evaluation left off.

Metric aggregation follows a four-level hierarchy: (1) macro-averaging across languages per task, with configurable aggregation modes for monolingual-only or cross-lingual grouping by input or output language; (2) macro-averaging across tasks per task group (e.g., all skill extraction tasks); (3) macro-averaging across task groups per task type (e.g., all ranking tasks); and (4) macro-averaging across task types for the overall benchmark score (e.g. ranking and classification). Structured output files include `results.json` for metrics, `checkpoint.json` for completion state, and `config.yaml` for configuration. The default ranking metrics are MAP, NDCG, R-Precision@10, and MRR, with additional metrics available such as Recall@K and Hit@K.

4 Broader Impact

4.1 A Community-Driven Ecosystem

As illustrated in Figure 1, WorkRB is sustained by three complementary pillars: industry partners contribute models and datasets for real-world tasks, academic labs contribute methodological and foundational innovations, and government institutions maintain the multilingual occupational ontologies (e.g., ESCO, O*NET) that form the taxonomic backbone of many tasks. This ecosystem is designed around a mutual incentive model: contributors gain visibility through citations and adoption of their resources, while in return they benefit from (i) scalable evaluation across tasks and languages with automated checkpointing, (ii) a standardized interface, (iii) discovery of state-of-the-art models and datasets, and (iv) the ability to extend the benchmark with proprietary tasks for internal use without disclosing sensitive data. To sustain growth, WorkRB provides clear contribution guidelines, issue-based task proposals, and continuous integration, with a community benchmark challenge planned at 2026 RecSys in HR workshop (pending acceptance).

4.2 Standardization towards Compliant & Representative Evaluation

Open standardization. Employment data is subject to increasingly strict regulation [9, 10, 16]; open-source benchmarks are well-positioned to support compliance. They enable independent verification, reproducible auditing, and transparent standardization, properties that proprietary benchmarks cannot offer. For this reason, the European Commission explicitly encourages the development of benchmarks and measurement methodologies in the context of accuracy and robustness requirements for high-risk AI systems. WorkRB embodies this principle by building its public tasks on openly licensed data, while its extensible design allows organizations to evaluate proprietary datasets and models internally under the same framework. Note, however, that legal compliance extends beyond benchmark accuracy. In the EU, AI providers must also address risks to fundamental rights such as dignity, autonomy, and non-discrimination. Such requirements are difficult to capture in computational metrics, making interdisciplinary assessment essential in practice. Public Employment Services (PES) show the practical relevance of this discussion. Through its 74 members across five regions, the World Association of PES (WAPES) provides a global platform for exchange on how employment services respond to labor market change, including AI. In this context, transparent, responsible, and context-sensitive evaluation approaches can support peer learning and promote the adoption of trustworthy AI in PES.

Multilinguality fosters representativeness. Work-domain AI is deployed worldwide, yet evaluation resources remain concentrated in high-resource languages. By centralizing multilingual evaluation across up to 28 languages through multilingual ontologies and diverse contributed datasets, WorkRB improves accessibility for underrepresented language populations, and shifts focus from English-only evaluations.

5 Conclusion & Future Work

We presented WorkRB, the first community-driven, open-source evaluation framework for AI in the work domain, addressing the fragmentation of evaluation in this multi-stakeholder domain. WorkRB unifies 13 recommendation and NLP tasks with dynamic multilingual ontology support, an extensible pip-installable toolkit, and a multi-stakeholder ecosystem bridging academia, industry, and public institutions. Future work includes extending task coverage to occupational activities, additional ontologies, and temporal evaluation dimensions, as well as incorporating career path recommendation and multi-branch architectures including large language models and re-ranking approaches. As the framework is designed to scale through continued community participation, we encourage contributions for new tasks, datasets, models, and metrics, as exemplified by the many contributions in this paper shaped by multiple stakeholders. Contributing is simple through available getting-started instructions and contribution guidelines at `https://github.com/techwolf-ai/WorkRB`.

References

[1] Aleksander Bielinski and David Brazier. 2025. 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 (RecSys-in-HR 2025), in conjunction with the 19th ACM Conference on Recommender Systems* (Prague, Czech Republic), Toine Bogers, Guillaume Bied, Jean-Joris Decorte, Chris Johnson, and Mesut Kaya (Eds.), Vol. 4046. CEUR, Article 5, 10 pages. https://ceur-ws.org/Vol-4046/RecSysHR2025-paper_5.pdf

[2] Matthias De Lange, Jens-Joris Decorte, and Jeroen Van Hautte. 2025. Unified Work Embeddings: Contrastive Learning of a Bidirectional Multi-task Ranker. *arXiv preprint arXiv:2511.07969* (2025).

[3] Jens-Joris Decorte, Matthias De Lange, and Jeroen Van Hautte. 2025. Multilingual JobBERT for Cross-Lingual Job Title Matching. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)* 4038 (2025). https://ceur-ws.org/Vol-4038/paper_367.pdf

[4] Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Design of negative sampling strategies for distantly supervised skill extraction. In *Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2022)* (Seatle, USA), Mesut Kaya, Toine Bogers, David Graus, Sepideh Mesbah, Chris Johnson, and Francisco Gutiérrez (Eds.), Vol. 3218. CEUR, Article 4, 7 pages. https://ceur-ws.org/Vol-3218/RecSysHR2022-paper_4.pdf

[5] Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester, and Chris Develder. 2021. JobBERT : understanding job titles through skills. In *FEAST, ECML-PKDD 2021 Workshop, Proceedings* (Online). 9. https://feast-ecmlpkdd.github.io/papers/FEAST2021_paper_6.pdf

[6] Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester, and Chris Develder. 2024. SkillMatch: Evaluating Self-supervised Learning of Skill Relatedness. *arXiv preprint arXiv:2410.05006* (2024).

[7] Jens-Joris Decorte, Jeroen Van Hautte, Chris Develder, and Thomas Demeester. 2025. Efficient Text Encoders for Labor Market Analysis. *arXiv preprint arXiv:2505.24640* (2025).

[8] Daniel Deniz, Federico Retyk, Laura García-Sardiña, Hermenegildo Fabregat, Luis Gasco, and Rabih Zbib. 2024. Combined Unsupervised and Contrastive Learning for Multilingual Job Recommendations. In *Proceedings of the 4th Workshop on Recommender Systems for Human Resources (RecSys in HR 2024), in conjunction with the 18th ACM Conference on Recommender Systems*.

[9] European Commission, Joint Research Centre. 2025. *AI Standards and Standardisation Landscape*. Technical Report. Publications Office of the European Union. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430

[10] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 16 (Jan. 2025), 54 pages. doi:10.1145/3696457

[11] Warren Jouanneau, Emma Jouffroy, and Marc Palyart. 2025. An Efficient Long-Context Ranking Architecture With Calibrated LLM Distillation: Application to Person-Job Fit. (2025).

[12] Warren Jouanneau, Marc Palyart, and Emma Jouffroy. 2024. Skill matching at scale: freelancer-project alignment for efficient multilingual candidate retrieval. (2024).

[13] Martin le Vrang, Agis Papantoniou, Erika Pauwels, Pieter Fannes, Dominique Vandensteen, and Johan De Smedt. 2014. Esco: Boosting job matching in europe with semantic interoperability. *Computer* 47, 10 (2014), 57–64.

[14] Magron, Antoine and Dai, Anna and Zhang, Mike and Montariol, Syrielle and Bosselut, Antoine. 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)*, Hruschka, Estevam and Lake, Thom and Otani, Naoki and Mitchell, Tom (Ed.). Association for Computational Linguistics, St. Julian's, Malta, 43–58. https://aclanthology.org/2024.nlp4hr-1.4/

[15] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2014–2037.

[16] New York City Council. 2023. Local Law 144: Automated Employment Decision Tools. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page. Effective July 5, 2023.

[17] O*NET Resource Center. n.d.. O*NET OnLine. https://www.onetonline.org/. Accessed: 2026-02-25.

[18] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[19] Federico Retyk, Luis Gasco, Casimiro Pio Carrino, Daniel Deniz, and Rabih Zbib. 2024. MELO: An Evaluation Benchmark for Multilingual Entity Linking of Occupations. In *Proceedings of the 4th Workshop on Recommender Systems for Human Resources (RecSys in HR 2024), in conjunction with the 18th ACM Conference on Recommender Systems*. https://recsyshr.aau.dk/wp-content/uploads/2024/10/RecSysHR2024-paper_2.pdf

[20] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Vol. 4. Now Publishers Inc.

[21] SkillsFuture Singapore. n.d.. Skills Framework. https://www.skillsfuture.gov.sg/skills-framework. Accessed: 2026-02-25.

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663* (2021).

[23] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32 (2019).

[24] Rabih Zbib, Lucas Alvarez Lacasa, Federico Retyk, Rus Poves, Juan Aizpuru, Hermenegildo Fabregat, Vaidotas Šimkus, and Emilia García-Casademont. 2022. Learning Job Titles Similarity from Noisy Skill Labels. *FEAST, ECML-PKDD 2022 Workshop* (2022). https://feast-ecmlpkdd.github.io/archive/2022/papers/FEAST2022_paper_4972.pdf

[25] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).