# A Recipe for Adapting Multilingual Embedders to OCR-Error Robustness and Historical Texts

**Andrianos Michail*, Stylianos Psychias*, Juri Opitz, and Simon Clematide**
Department of Computational Linguistics, University of Zurich, Switzerland
{andrianos.michail,stylianos.psychias,juri.opitz,simon.clematide}@cl.uzh.ch

**Abstract**

Modern multilingual text embedding models excel at semantic search on contemporary text but their performance degrades measurably on digitized historical documents. This issue is especially pronounced for underrepresented languages such as Luxembourgish, where historical materials combine evolving spelling conventions with OCR artifacts absent from standard training data. To address these challenges, we introduce OCR M-GTE, a pair of multilingual embedding models adapted for OCR robustness and historical texts, and show that the observed degradation can be mitigated through a simple multi-step training procedure tailored to historical variants and OCR noise. We evaluate the models on standard semantic search tasks, simulated OCR degradation, and genuine historical collections, observing consistent improvements under OCR-induced noise and on genuine historical data while maintaining comparable performance on clean modern text. Our ablation findings suggest that multilingual embedding models can be effectively adapted to perform robust cross-lingual search in heterogeneous European digitized corpora. We release our adapted models, code, and datasets under the AGPL-3.0 license: https://github.com/impresso/ocr-robust-multilingual-embeddings

**Keywords:** multilingual embedding models, OCR-noise-robust embeddings, cross-lingual semantic search

## 1. Introduction

The large-scale digitization of historical texts has fundamentally expanded access to archival and cultural heritage materials. In this context, semantic search based on text embeddings has emerged as a powerful alternative to traditional keyword-based retrieval (Karpukhin et al., 2020). Modern multilingual embedding models, trained on clean contemporary text, perform strongly on multilingual and cross-lingual benchmarks (Enevoldsen et al., 2025) and now underpin many retrieval-augmented generation (RAG) systems (Lewis et al., 2020). By mapping queries and documents to dense vector representations, they enable contextually relevant retrieval and improve the accuracy and factuality of RAG outputs.

However, the effectiveness of these dense representations degrades on OCR-affected text. Errors arising from poor scans, historical fonts, and document decay disrupt tokenization and sentence integrity, posing a distinct challenge for lexical representations (Taghva et al., 1996; Mutuvi et al., 2018). Because such noise is largely absent from model pretraining, off-the-shelf embeddings generalize poorly to historical digitized texts, exhibiting substantial drops in topic modeling (Zosa et al., 2021) and cross-lingual semantic search performance (Michail et al., 2025b). This degradation has motivated efforts to post-train embedding models so that they become more resilient to character-level noise.

Complementary research has explored adaptation strategies for improving robustness to OCR-induced perturbations, inspired by earlier work on typo-tolerant dense representations (Tasawong et al., 2023; Sidiropoulos and Kanoulas, 2022). In this line of work, termed *noise adaptation*, the encoder is fine-tuned on clean—noised text pairs generated via stochastic character-level edits (substitution, insertion, deletion) applied to approximately 5% of the characters. Controlled experiments demonstrate significant robustness gains on OCR-like noise while maintaining comparable performance on clean-text evaluations (Michail et al., 2025b). While such denoising methods improve general robustness, language-specific challenges remain, particularly for underrepresented languages.

A particularly illustrative case is Luxembourgish, a language with limited representation in multilingual embedding pretraining (Lothritz et al., 2022). Historical Luxembourgish texts spanning 150 years combine evolving orthography and OCR-induced artifacts (Michail et al., 2025c). Previous work has paired historical Luxembourgish with large language model (LLM) translations into German, French, and English to provide training signals that improve model understanding of historical Luxembourgish. These training regimes have been shown to improve performance in cross-lingual semantic search on historical data (Michail et al., 2025c). This line of research highlights how historical language variation compounds OCR-induced degradation, underscoring the need for multilingual models that are robust to both sources of noise.

In this work, we unify and extend these two complementary approaches to support semantic

* Equal contribution.



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 929–935
11–16 May 2026. 

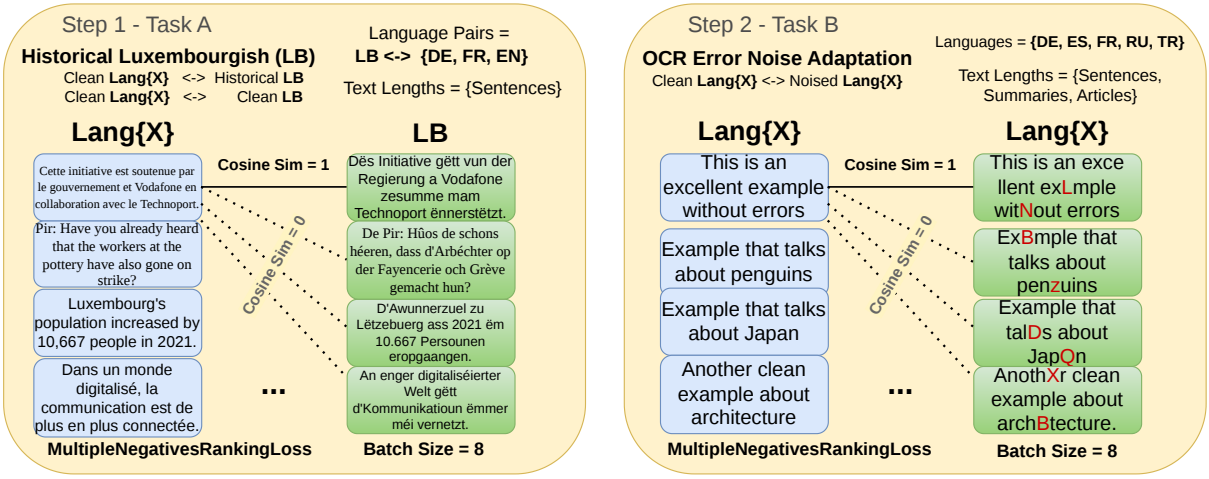


Figure 1: Overview of the two-step adaptation procedure. Step 1 (Task A) aligns historical Luxembourgish with contemporary European languages using cross-lingual pairs. Step 2 (Task B) increases robustness to OCR-like noise through monolingual denoising training across multiple languages and text lengths.

search across diverse European languages and text lengths. The result is an embedding model specifically adapted for digitized historical collections. Concretely, we adapt the multilingual long-context embedding model M-GTE (Zhang et al., 2024) through a two-step procedure: (i) **Task A – Historical Luxembourgish**, using cross-lingual pairs between historical Luxembourgish and contemporary European languages, and (ii) **Task B – OCR Error Noise Adaptation**, a monolingual fine-tuning stage across six European languages that increases tolerance to OCR-like noise at different text lengths. Figure 1 provides an overview of our proposed two-step adaptation procedure.

We present **OCR M-GTE**, a family of multilingual embedding models adapted for OCR-error-affected historical text.

Our main contributions are as follows:

1. We conduct an ablation study over variants of a two-step adaptation strategy—historical Luxembourgish language acquisition (cross-lingual pairs with DE/FR/EN) followed by character-level OCR-error noise adaptation—demonstrating improved robustness to noisy inputs.

2. We release two multilingual embedding models adapted for historical and noisy European texts: (i) the denoising-trained generalist model, *gte-multilingual-base-ocr-noise-robust*, and (ii) the historical Luxembourgish specialist model, *gte-multilingual-base-histlux-ocr-noise-robust*, which trades slight degradation on modern text for improved performance on historical Luxembourgish.

## 2. Method

### 2.1. Base Model — Multilingual GTE

Similar to previous studies (Chari et al., 2025; Ezerceli et al., 2025; Zhan et al., 2025), we adopt the *Multilingual General Text Embedding* model (M-GTE) (Zhang et al., 2024) as our base model. M-GTE supports multilingual and long-context text representations and is trained using contrastive learning. Its pretraining corpus already includes 50,000 Luxembourgish sentence pairs, and the final training stage further refines the representations using a hard-negative contrastive loss. Owing to its broad language coverage, moderate parameter count, support for extended context lengths, and diverse training objectives, M-GTE provides a reliable foundation for developing an OCR-robust embedding model.

### 2.2. Historical Luxembourgish (Task A)

We follow the "Mix-it-all" fine-tuning configuration for historical and contemporary Luxembourgish language acquisition proposed by Michail et al. (2025c). This setup combines 20,000 historical Luxembourgish sentences paired with machine-translated contemporary German, French, and English counterparts, yielding 60,000 historical Luxembourgish cross-lingual pairs. To stabilize training and enhance alignment with contemporary usage, the mixture additionally includes modern Luxembourgish—French (40K) and Luxembourgish–English (20K) pairs drawn from the training data of the contemporary Luxembourgish specialist *Luxembedder* model (Philippy et al., 2025).

The model is then fine-tuned on the resulting 120k pairs using the Multiple Negatives Ranking

| Source | Type | Avg. Length (tokens) | Languages | Total Pairs |
|---|---|---|---|---|
| **Historical Luxembourgish (Task A)** | | | | |
| Historical Luxembourgish | Cross-lingual | 13 | LB ↔ DE/FR/EN | 60,000 |
| Modern Luxembourgish | Cross-lingual | 18 | LB ↔ FR/EN | 60,000 |
| **OCR-Error Noise Adaptation (Task B)** | | | | |
| TED | Monolingual | 16 | DE, FR | 20,000 |
| Historical Articles | Monolingual | 315 | DE, FR | 20,000 |
| MLSum Articles | Monolingual | 550 | DE, ES, FR, RU, TR | 10,000 |
| MLSum Summaries | Monolingual | 20 | DE, ES, FR, RU, TR | 10,000 |
| **Total** | | | | 180,000 |

Table 1: Datasets used for adapting the embedding model to Luxembourgish and OCR-error noise.

Loss (Henderson et al., 2017) from *SentenceTransformers* (Reimers and Gurevych, 2019), with a batch size of 8.

### 2.3. OCR-Error Noise Adaptation (Task B)

We follow the monolingual denoising procedure introduced by Michail et al. (2025b). Random character-level perturbations are applied to 5% of the characters in each corpus to generate noisy–clean sentence pairs for fine-tuning. This procedure is applied to the following corpora:

**TED:** 10k German and 10k French denoising sentence pairs (Michail et al., 2025b) derived from the TED corpus (Qi et al., 2018).

**Historical Articles:** High-quality German and French newspaper articles sourced from a European digital archive. During sampling, we required at least 98% of all tokens to appear in the respective language's vocabulary, resulting in clean text with minimal residual OCR errors.

**MLSum:** News articles and their summaries (Scialom et al., 2020) covering five languages—German (DE), Spanish (ES), French (FR), Russian (RU), and Turkish (TR). Each language subset includes 2,000 records. Article texts and reference summaries are treated as two independent monolingual sources (average length: 20 tokens for summaries, 550 for articles).

The model is further fine-tuned on up to 60k pairs using the Multiple Negatives Ranking Loss (Henderson et al., 2017) objective from *SentenceTransformers*, with a batch size of 8. Each batch contains pairs from a single language and source corpus.

### 2.4. Overview of Fine-tuning Datasets

Table 1 summarizes the datasets used for both adaptation tasks: **Historical Luxembourgish** and **OCR-Error Noise Adaptation**. The table lists the source, average text length, language coverage, and number of training pairs. The Luxembourgish component involves cross-lingual sentence pairs, divided into historical and contemporary subsets, while the denoising component consists of monolingual perturbation-based data. Together, these resources constitute the complete adaptation setup for developing OCR-robust multilingual embeddings.

### 2.5. Two-step Fine-tuning

Preliminary experiments showed that jointly optimizing the two objectives in a single training run led to unstable results. We therefore adopt a sequential two-step fine-tuning procedure. In Step 1 (Task A), the model learns to represent the underrepresented historical variant of Luxembourgish, including its orthographic variability across 150 years of texts that contain occasional OCR errors. In Step 2 (Task B), the model is further adapted to stochastic character-level perturbations that emulate OCR errors across six European languages spanning three writing systems.

## 3. Evaluation Methods

We evaluate the effectiveness of our adaptation method on cross-lingual datasets that contain both realistically simulated and naturally occurring OCR noise. To verify that OCR-specific fine-tuning does not degrade performance on standard semantic search data, we evaluate the models on clean, noise-free semantic ranking control tasks.

Specifically, we test three complementary conditions: (i) **Control**: Clean cross-lingual semantic ranking tasks; (ii) **Realistic**: OCR-degraded versions of the clean control evaluation sets; and (iii) **Organic**: a historical Luxembourgish bitext-mining task containing real OCR errors. We first evaluate on clean control benchmarks to establish baseline performance.

### 3.1. Clean Control Tasks

**CLSD:** We evaluate on the German–French test set of Cross-Lingual Semantic Discrimination (CLSD)

| Model | HistLUX | CLSD — Blackletter/SN | | | CLSD — Salt & Pepper Noise | | | CLEAN CLSD | X-STS | Averages | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | LB↔DE/FR/EN (Avg) | Clean → BL/SN | BL/SN → BL/SN | DE↔FR (Avg) | Clean → SnP | SnP → SnP | DE↔FR (Avg) | DE↔FR (Avg) | EN-TR/ES/AR (Avg) | OCR Tasks | Clean Tasks |
| **Multilingual-GTE Base** | 83.78 | 79.10 | 77.20 | 78.15 | 81.32 | 81.92 | 81.62 | 90.50 | 79.78 | 81.18 | 85.14 |
| **One-step: OCR-Error Noise Adaptation only (Task B)** | | | | | | | | | | | |
| + TED | 87.44 | 82.26 | 79.92 | 81.09 | 82.79 | 83.77 | 83.28 | 92.84 | 78.70 | 83.94 | 85.77 |
| + Historical Articles | 87.37 | 82.15 | 79.75 | 80.95 | 82.91 | 83.97 | 83.44 | 92.83 | 79.08 | 83.92 | 85.96 |
| + MLSum Articles | 87.49 | 82.24 | 79.84 | 81.04 | 83.04 | 83.96 | 83.50 | 92.92 | 79.12 | 84.01 | 86.02 |
| + MLSum Summaries | 87.43 | 82.60 | 79.98 | 81.29 | **83.66** | **84.39** | **84.03** | **93.26** | **79.17** | 84.25 | **86.22** |
| **Two-step: OCR-Error Noise Adaptation (Task B) followed by Historical Luxembourgish (Task A)** | | | | | | | | | | | |
| + TED | 96.43 | 78.61 | 76.58 | 77.60 | 72.65 | 74.39 | 73.52 | 91.38 | 63.30 | 82.52 | 77.34 |
| + Historical Articles | 96.46 | 78.24 | 76.46 | 77.35 | 72.75 | 74.71 | 73.73 | 91.12 | 64.69 | 82.51 | 77.90 |
| + MLSum Articles | 96.46 | 78.32 | 76.75 | 77.54 | 72.56 | 74.71 | 73.64 | 91.39 | 64.74 | 82.54 | 78.07 |
| + MLSum Summaries | 96.49 | 77.93 | 76.26 | 77.10 | 72.46 | 74.29 | 73.38 | 91.06 | 63.95 | 82.32 | 77.51 |
| **Two-step: Historical Luxembourgish (Task A) followed by OCR-Error Noise Adaptation (Task B)** | | | | | | | | | | | |
| LB only | 97.28 | 78.57 | 76.64 | 77.61 | 70.75 | 72.51 | 71.63 | 91.16 | 58.68 | 82.17 | 74.92 |
| + TED | 97.43 | 81.90 | 80.12 | 81.01 | 76.05 | 76.96 | 76.51 | 92.35 | 69.91 | 84.98 | 81.13 |
| + Historical Articles | 97.48 | 81.63 | 79.68 | 80.66 | 75.87 | 76.79 | 76.33 | 92.26 | 69.74 | 84.82 | 81.00 |
| + MLSum Articles | 97.52 | 81.83 | 79.78 | 80.81 | 75.97 | 76.92 | 76.45 | 92.37 | 70.12 | 84.92 | 81.24 |
| + MLSum Summaries | **97.59** | **82.60** | **80.46** | **81.53** | 76.34 | 77.14 | 76.74 | 92.50 | 71.16 | **85.29** | 81.83 |

Table 2: Average results across five fine-tuning seeds. Metrics: **Precision@1** for CLSD and HistLUX, **Spearman correlation (×100)** for X-STS. OCR-Error Noise Adaptation data sources are additive within each subheader. Dotted lines indicate models included in the public release.

introduced by Michail et al. (2025a). The task assesses whether multilingual embedding models rank the correct parallel sentence highest among four semantically similar distractors. Performance is reported as Precision@1. The benchmark comprises two datasets—WMT19 (Barrault et al., 2019) and WMT21 (Akhbardeh et al., 2021)—for each translation direction (DE→FR, FR→DE), and the results are averaged across the four configurations.

**X-STS:** We further assess cross-lingual semantic textual similarity using the SemEval-2017 Task 1 (X-STS) benchmark (Cer et al., 2017). Evaluations are performed on languages not included in fine-tuning: EN↔TR, EN↔ES, and EN↔AR. Scores are reported as Spearman correlation ($\times 100$) between model-predicted and human-rated similarities.

### 3.2. Realistic OCR Noise CLSD

Building on the clean **CLSD** dataset described above, we assess model robustness under realistic OCR-induced degradation using an extended version of the dataset. The digitization process follows the three-step pipeline of Michail et al. (2025b): (S1) print the text under realistic conditions, (S2) apply visual degradations, and (S3) re-digitize the output using an OCR engine.

We consider two noise types: (1) **BlackLetter (BL) / Scanned Noise (SN)**: distortions from historical typefaces and low-quality scans; and (2) **Salt-and-Pepper (SnP)**: artifacts from paper degradation and aged-document backgrounds.

Two evaluation configurations are considered. In the first, clean source-language queries are matched against OCR-degraded target texts, simulating user search in digitized archives. In the second, both source and target texts contain OCR noise, simulating recommendation or cross-lingual text-reuse tasks within historical corpora. For each noise type and configuration—four evaluations in total—we report mean Precision@1 averaged over both language directions and source datasets.

### 3.3. Historical Luxembourgish Bitext-Mining

We further evaluate the adapted models on the **Historical Bitext Mining** test set introduced by Michail et al. (2025c). The dataset comprises 233 digitized historical Luxembourgish newspaper articles (1840–1950), sentence-segmented and machine-translated into modern German, French, and English. The task requires the model to rank the correct translations highest among alternative candidates in texts exhibiting naturally occurring OCR errors. Performance is reported as the average Precision@1 across the three language pairs.

## 4. Results

We conduct a fine-tuning ablation study to evaluate the impact of data scale and the task order. The ablation varies the amount of data used in Task B and compares three configurations: (1) One-step: Task B only (noise adaptation only); (2) Two-step: Task B followed by Task A; and (3) Two-step: Task A followed by Task B. Table 2 summarizes the results.

Overall, Task B fine-tuning consistently improves performance across all evaluation settings, confirming the findings of Michail et al. (2025b). Incorporating additional heterogeneous data sources yields small but steady gains.

The effect of task order is pronounced: performing Task A before Task B results in substantially higher performance on both OCR-degraded and clean-text tasks than the reverse order. This configuration achieves the highest average scores on

OCR-related evaluations, with only a minor trade-off on the X-STS benchmark.

Interestingly, Task A—based on historical texts that include naturally occurring OCR misrecognitions and an elevated punctuation rate of 6.8% (compared to 3.6% in clean text)—slightly reduces performance on the CLSD task with Salt-and-Pepper noise, which similarly introduces spurious punctuation (5.9% of characters). In contrast, performance on CLSD with BlackLetter/Scanned-Noise, which does not add punctuation artifacts, remains largely unaffected.

For general multilingual applications that require robustness without any loss in performance on X-STS, we release the model *gte-multilingual-base-ocr-noise-robust*, fine-tuned solely on the Task B (OCR-error noise adaptation) objective using the complete heterogeneous data mix. For European-language and historical Luxembourgish use cases, we release the model *gte-multilingual-base-histlux-ocr-noise-robust*, trained with the two-step configuration—Task A followed by Task B—on the full dataset mix. As a side effect of the historical Luxembourgish adaptation, this model exhibits an average decrease of 8 Spearman points on the X-STS benchmark.

## 5. Conclusion

We present **OCR M-GTE**, two multilingual embedding models adapted for robust semantic search in OCR-affected and historical texts. Our two-step adaptation strategy—combining parallel historical Luxembourgish data with multilingual OCR denoising—effectively mitigates the degradation observed in off-the-shelf models on digitized corpora.

Experimental results show robust gains under realistic OCR noise and on naturally degraded historical Luxembourgish data, while preserving strong performance on clean-text evaluations. The sequential adaptation order—historical Luxembourgish language acquisition followed by OCR-error noise adaptation—yields the best balance between OCR robustness and preservation of modern text performance.

We release two models on Hugging Face to support downstream research and real-world applications: (i) a denoising-trained generalist for multilingual retrieval, and (ii) a Luxembourgish specialist optimized for OCR-rich archives. These resources advance the development of reliable semantic search systems that improve access to and exploration of digitized historical multilingual collections.

## 6. Limitations

Our work builds on existing research and uses a broad evaluation suite that includes both historical and contemporary datasets. However, generalization of the observed improvements to other data is not guaranteed. For example, how well these models that were originally excellent in Information Retrieval have preserved this capabilities is unknown and we invite future research examines this both in the historic and in the contemporary domain. The adaptation regime is simple and intuitive, and while we tested multiple configurations across five fine-tuning seeds, it remains unclear whether this approach represents the most effective solution.

## 7. Bibliographical References


Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity — multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages

1–14, Vancouver, Canada. Association for Computational Linguistics.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2025. Improving low-resource retrieval effectiveness using zero-shot linguistic similarity transfer. In *Advances in Information Retrieval*, pages 290–306, Cham. Springer Nature Switzerland.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Özay Ezerceli, Gizem Gümüşçekiçci, Tuğba Erkoç, and Berke Özenç. 2025. Turkembed4retrieval: Turkish embedding model for retrieval task. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems, volume 33 (NeurIPS 2020)*, pages 9459–9474.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Andrianos Michail, Simon Clematide, and Rico Sennrich. 2025a. Examining multilingual embedding models cross-lingually through llm-generated adversarial examples. ArXiv preprint.

Andrianos Michail, Juri Opitz, Yining Wang, Robin Meister, Rico Sennrich, and Simon Clematide. 2025b. Cheap character noise for OCR-robust multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11705–11716, Vienna, Austria. Association for Computational Linguistics.

Andrianos Michail, Corina Raclé, Juri Opitz, and Simon Clematide. 2025c. Adapting multilingual embedding models to historical Luxembourgish. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 291–298, Albuquerque, New Mexico. Association for Computational Linguistics.

Stephen Mutuvi, Antoine Doucet, Moses Odeo, and Adam Jatowt. 2018. Evaluating the impact of ocr errors on topic modeling. In *Maturity and Innovation in Digital Libraries*, pages 3–14, Cham. Springer International Publishing.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders

for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2132–2136, New York, NY, USA. Association for Computing Machinery.

Kazem Taghva, Julie Borsack, and Allen Condit. 1996. Evaluation of model-based retrieval effectiveness with ocr text. *ACM Trans. Inf. Syst.*, 14(1):64–93.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115, Toronto, Canada. Association for Computational Linguistics.

Shaoxiong Zhan, Hai Lin, Hongming Tan, Xiaodong Cai, Hai-Tao Zheng, Xin Su, Zifei Shan, Ruitong Liu, and Hong-Gee Kim. 2025. Lexsembridge: Fine-grained dense representation enhancement through token-aware embedding augmentation.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Elaine Zosa, Stephen Mutuvi, Mark Granroth-Wilding, and Antoine Doucet. 2021. Evaluating the robustness of embedding-based topic models to ocr noise. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings*, page 392–400, Berlin, Heidelberg. Springer-Verlag.