

LëtZCross: A Cross-Lingual Page-Level Benchmark for Multimodal Retrieval over Luxembourgish Documents

Omar El Bachyr^{1,*} Fred Philippy¹ Laura Maria Bernardy¹
Saad Ezzini² Jacques Klein¹ Tegawendé F. Bissyandé¹

¹University of Luxembourg, Luxembourg

²King Fahd University of Petroleum and Minerals, Saudi Arabia

*Correspondence: omar.elbachyr@uni.lu

Abstract

Recent page-image retrievers such as ColPali (Faysse et al., 2025) have improved retrieval over visually rich documents, yet little is known about how they behave in cross-lingual, low-resource settings. We introduce *LëtZCross*, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF documents, with document pages indexed as images and queries provided in English, French, German, and Luxembourgish. The benchmark combines text-focused QA pairs with visually grounded QA pairs, covering both textual and visual retrieval needs in PDF-based RAG. We use *LëtZCross* to compare OCR-based text-only retrievers with ColPali-style page-image retrievers and find that the latter perform better across query languages in this system-level comparison. We also examine single-language and multilingual fine-tuning. Fine-tuning transfers across query languages, with French yielding the highest mean performance on Luxembourgish queries among the single-language settings. In the multilingual setting, including Luxembourgish gives the strongest results and substantially improves retrieval for Luxembourgish queries.

1 Introduction

Retrieval-augmented generation (RAG) and document retrieval systems increasingly operate in multilingual environments, where user queries and source documents may be written in different languages. In this setting, cross-lingual information retrieval (CLIR) has advanced substantially with dense retrieval architectures and multilingual encoders (Karpukhin et al., 2020; Conneau et al., 2020). However, most current evaluations focus on text-centric, high-resource scenarios and do not reflect the challenges of visually rich PDF documents.

Recent vision–language retrieval models, such

as DSE (Ma et al., 2024) and ColPali (Faysse et al., 2025), improve retrieval over document pages by jointly modeling textual and visual evidence. Yet their behavior in low-resource language settings remains underexplored. We study this challenge through the case of Luxembourgish, a low-resource language for which NLP resources are growing but still limited compared to high-resource languages (Lothritz et al., 2022; Bernardy, 2022; Plum et al., 2025).

To address this gap, we introduce *LëtZCross*, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF pages, where pages are represented as images. The benchmark combines text-focused and visually grounded QA pairs, and is designed around realistic page-level retrieval needs in RAG pipelines, where answers may depend on textual content, visual structure, or both, and where queries may be issued in multiple languages.

Our main contributions are:

- We introduce *LëtZCross*, a benchmark for cross-lingual page-level retrieval over Luxembourgish PDF pages, where pages are represented as images.
- We benchmark multilingual text-only retrievers and late-interaction page-image retrievers on this task.
- We study query-language-specific fine-tuning for cross-lingual retrieval over Luxembourgish document pages.
- We analyze multilingual fine-tuning settings for retrieval over Luxembourgish document pages as a low-resource language use case.

Overall, our findings show that late-interaction page-image retrievers are strong candidates for cross-lingual retrieval over Luxembourgish document pages, outperforming the evaluated OCR-

based text-only baselines in a system-level comparison. We further show that query-language-specific fine-tuning transfers across languages, while multilingual fine-tuning is most effective when it includes the target low-resource language. Together, these results position *LëtZCross* as a useful benchmark for studying cross-lingual page-level retrieval over Luxembourgish PDF documents as a low-resource language in RAG settings.

We release *LëtZCross* and all experimental code and resources through our [GitHub repository](#).

2 Related Work

Luxembourgish NLP-Resources The Luxembourgish NLP landscape is still under development. Natural Language Processing techniques for Luxembourgish are primarily developed within a low-resource setting and rely heavily on manual efforts, such as data collection and annotation. In recent years, this approach has enabled the development of several foundational resources and applications, including sentiment analysis systems (Sirajzade et al., 2020; Gierschek, 2022), a Luxembourgish dependency parsing treebank (Plum et al., 2024), and zero-shot topic classification (Philippy et al., 2024). These initiatives have contributed to a better understanding of the linguistic properties of Luxembourgish through computational methods. Within this context, Lothritz et al. introduced LuxemBERT (Lothritz et al., 2022), a language model trained using artificial data augmentation strategies. In addition to the model itself, they expanded Luxembourgish resources for several downstream tasks, including part-of-speech tagging, named entity recognition, and news classification. Two further language models have since been established for Luxembourgish: LuXGPT (Bernardy, 2022), developed using transfer learning techniques, and, more recently, LUXT5 (Plum et al., 2025), which benefits from multilingual pretraining. Beyond language modeling, practical NLP applications have also emerged. These include LUX-ASR (Gilles et al., 2023) for automatic speech recognition, as well as systems for automatic comment moderation (Ranasinghe et al., 2023) and orthographic correction (Purschke, 2020).

Cross-lingual retrieval Transformer-based Cross-Lingual Information Retrieval (CLIR) builds on multilingual representation learning and dense retrieval architectures that enable semantic matching across languages. Early foundations

include BERT (Devlin et al., 2019), which introduced contextual transformer representations for retrieval models, and XLM-R (Conneau et al., 2020), a multilingual transformer that provides strong cross-lingual embeddings. Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) later established the dual-encoder architecture widely used in dense retrieval systems. Building on these foundations, more recent work explores unsupervised and generative approaches, such as Unsupervised Multilingual Dense Retrieval (UMR) proposed by (Huang et al., 2024), which uses generative pseudo-labeling to train dense retrievers without parallel supervision, and research on adapting LLMs for retrieval tasks (Guo et al., 2024). Recent multilingual embedding models such as BGE-M3 (Chen et al., 2024), Qwen3 Embedding (Zhang et al., 2025), and voyage-4-nano (AI, 2026) further improve cross-lingual semantic representations and are widely used as baselines in modern CLIR systems.

Vision Language Retrieval Text-only encoders have achieved strong performance in information retrieval (IR), with models such as (Chen et al., 2024; Zhang et al., 2025; AI, 2026) demonstrating high effectiveness across multilingual and dense retrieval tasks. However, recent studies in IR have advanced beyond conventional text-only systems by incorporating visual features. DSE (Ma et al., 2024) propose the use of Vision-Language Models (VLMs) to encode full document screenshots, successfully preserving both text and visual information without the need for optical character recognition (OCR). Similarly, ColPali (Faysse et al., 2025) addresses the shortcomings of text-focused retrieval in the context of visually rich documents by adapting ColBERT (Khattab and Zaharia, 2020) late-interaction similarity scoring mechanism to work with page images. This produces multi-vector visual embeddings that enable fine-grained matching between query tokens and specific visual areas on document pages.

3 LëtZCross Benchmark Construction

LëtZCross is a cross-lingual page-level retrieval benchmark for Question-Answering (QA), built from Luxembourgish documents, with queries in English (EN), French (FR), German (DE), and Luxembourgish (LB). Documents are represented as page images, and answers are grounded in the corresponding pages. Figure 1 summarizes the process,

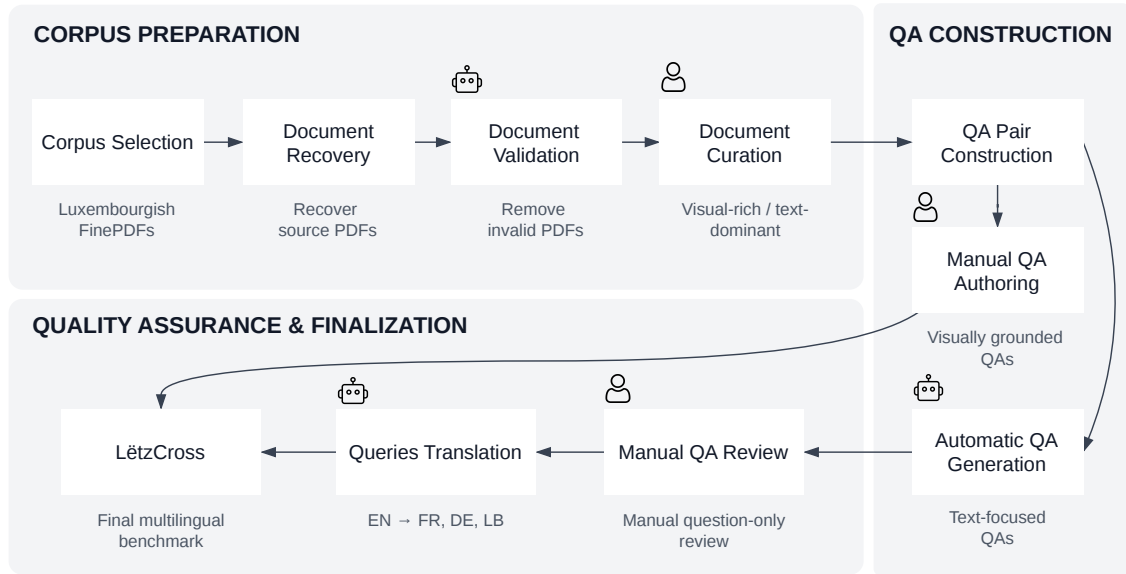


Figure 1: Overview of the LëtZCross benchmark construction pipeline.

with steps described below.

3.1 PDF Documents Acquisition

We obtain the original PDF documents by iterating over the Luxembourgish (LB) split of FinePDFs (Kydliček et al., 2025) dataset and retrieving each document from its source. When documents are archived in Common Crawl, we extract the corresponding PDF directly from the web archive; otherwise, we download the PDF from the original URL provided in the dataset. This process allows us to recover the original PDF files associated with the dataset entries.

3.2 Filtering and Curation of PDF Documents

After downloading the PDFs, we perform an automatic validation step to remove corrupted or unreadable documents by verifying that each PDF can be successfully parsed and processed. We then manually classify the validated PDFs into visually rich and text-dominant subsets. A document is considered visually rich when it contains informative visual elements, such as charts, diagrams, or tables, that convey substantive information beyond the surrounding text. Non-informative visuals, such as portraits, landscapes, and logos, are disregarded. Documents whose information is conveyed primarily through continuous text, with few or no informative visual elements, are classified as text-dominant. The resulting corpus serves as the foundation for the subsequent Question–Answer generation step.

3.3 Question–Answer Generation

We then use the validated PDFs from the previous step to generate Question–Answer (QA) pairs. We generate queries in English to ensure high-quality and consistent question formulation, as this yields more reliable QA pairs on Luxembourgish pages. The benchmark combines two complementary QA sources: manually authored visually grounded QA pairs for visually rich pages, and automatically generated text-focused QA pairs for text-dominant pages.

Manual visually grounded QA authoring. For the subset of PDFs containing rich visual components, a native Luxembourgish speaker manually crafts page-level QA pairs that intentionally target visually grounded evidence, such as charts, plots, diagrams, figures, and other layout-dependent elements, rather than relying solely on textual content. Questions are phrased to be self-contained and answerable from a single page image, with concise answers grounded in the visual content.

Automatic text-focused QA generation. Using the text-dominant PDF subset, we automatically generate page-level QA pairs with GPT-5-mini. Each page is processed independently as an image and prompted to produce self-contained questions with concise answers grounded in its textual content. To validate the generated QA pairs, each query and its corresponding PDF are provided to Gemini-2.5-Flash, which extracts an answer from

the document. We then compare the extracted answer with the *gold answer* using both exact and semantic matching; the latter is performed by prompting GPT-5-mini to assess whether the two answers convey the same meaning. Overall, 90.28% of the QA pairs were judged semantically equivalent by GPT-5-mini, and 71.50% contained the exact literal gold answer. QA pairs that do not yield consistent answers are discarded, resulting in a consistency-filtered set for the subsequent manual validation step.

3.4 Quality Control and RAG Alignment

As a follow-up step, the manually authored QA pairs are included directly in the final benchmark. In contrast, the automatically generated QA pairs undergo a manual validation phase focused exclusively on question formulation: annotators are shown only the questions and revise them when needed to produce clear, natural queries that reflect RAG-relevant information needs (e.g., realistic, retrieval-oriented queries).

3.5 Cross-Lingual Query Translation

To enable cross-lingual retrieval evaluation, we translate the finalized English queries into German, French, and Luxembourgish using a controlled LLM prompt that enforces semantic preservation and prevents paraphrasing or additional explanations. This produces aligned query sets across languages, where each translated query is intended to preserve the same retrieval intent as the original English query. We use Gemini-3-Flash for German and French, and Gemini-3-Pro for Luxembourgish. This choice was made after manually inspecting a small set of trial translations: Gemini-3-Pro produced better Luxembourgish translations, while Gemini-3-Flash was sufficient for German and French.

The final benchmark contains 579 QA pairs over 908 document pages. Table 1 summarizes the benchmark statistics, including the distribution between automatically generated text-focused QA pairs and manually authored visually grounded QA pairs. All prompts used during benchmark construction are available in the accompanying repository.

4 Methods

This section presents our method for investigating cross-lingual retrieval over Luxembourgish doc-

Benchmark statistic	Value
Number of pages	908
Number of queries	579
Avg. query length	19.5
Avg. corpus item length	722.5
QA type distribution	
Auto-generated text-focused	530
Manual visually grounded	49

Table 1: Overview statistics of the *LëtZCross* evaluation benchmark. Query and corpus item lengths are reported in tokens.

ument pages as a low-resource case study using late-interaction page-image retrievers.

4.1 Training Data Construction

We construct the training data by scraping Luxembourgish Wikipedia articles and converting them into PDF documents. We first apply size-based filtering with minimum and maximum page thresholds to exclude trivial or overly long documents (e.g., pages containing only footnotes). QA pairs are then generated from 4,977 PDFs with 11,833 pages using the same automatic text-focused QA generation procedure as in the benchmark (Section 3.3). This resulted in 22,028 QA pairs. In addition, we construct 450 QA pairs from the same Wikipedia-derived pipeline and reserve them as a held-out development split for fine-tuning. To extend the training set for cross-lingual retrieval, we translate queries into German and French, and a 5,000-query subset into Luxembourgish, following the same query translation procedure as in the benchmark (Section 3.5).

4.2 Retriever and Training Objective

We fine-tune a late-interaction page-image retriever based on ColPali (Faysse et al., 2025) using the training setup provided by the underlying framework. Each document page is treated as an image and split into visual patches, which are encoded into contextualized image patch embeddings, while the query text is tokenized and encoded into query token embeddings. Both are represented in a shared latent space, and relevance is computed through late interaction between query tokens and page patches.

Let $q_i = \{q_{i1}, \dots, q_{i|q_i|}\}$ denote the query token embeddings of query i , and let $d_j = \{d_{j1}, \dots, d_{j|d_j|}\}$ denote the document image patch embeddings of page j . The score between query i

and page j is:

$$\text{score}(q_i, d_j) = \sum_{t=1}^{|q_i|} \max_{u=1, \dots, |d_j|} (q_{it}^\top d_{ju}) \quad (1)$$

We use the original ColBERT (Khattab and Zaharia, 2020) contrastive objective with in-batch negatives:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{score}(q_i, d_i)/\tau)}{\sum_{j=1}^B \exp(\text{score}(q_i, d_j)/\tau)} \quad (2)$$

where τ is a temperature hyperparameter. Each training instance consists of a query paired with its corresponding document page image as the positive example, while the remaining $B - 1$ pages in the batch serve as in-batch negatives. We do not use external hard negatives or additional negative sampling.

4.3 Fine-Tuning Configurations

We consider two fine-tuning settings. In the *per-language* setting, the retriever is fine-tuned separately on English, French, and German queries. In the *multilingual* setting, a single model is trained on mixed-language data, using either EN+FR+DE or EN+FR+DE+LB. In all cases, document pages remain in Luxembourgish, and only the query language varies. To analyze visual adaptation, we also compare the default setting, which updates both textual and visual components, with a text-component-only variant. The EN, FR, and DE settings each use 22,028/450 train/dev QA pairs, while EN+FR+DE and EN+FR+DE+LB use 21,000/1,350 and 20,000/1,800 train/dev samples, respectively.

5 Experimental Setup

Models. Based on the input type, we consider two families of retrievers in our experiments: text-only retrievers and late-interaction page-image retrievers. The text-only baselines are bge-m3 (Chen et al., 2024), Qwen3-Embedding-0.6B (Zhang et al., 2025), multilingual-e5-large (Wang et al., 2024), and jina-embeddings-v4 (Günther et al., 2025)¹. The late-interaction page-image retrievers are colSmol-256M, colSmol-500M, colpali-v1.3, and colqwen2.5-v0.2 (Faysse et al., 2025), and colnomic-embed-multimodal-3b (Nomic AI, 2025).

¹Although jina-embeddings-v4 is a multimodal embedding model, we use it here in a text-only setting.

Text Extraction Pipeline. The text-only retrievers described above use page-level text extracted with unstructured[pdf] v0.18.15 under the hires strategy, combining Poppler page rendering, YOLOX layout detection, Luxembourgish Tesseract OCR (ltz), and table-structure extraction. Existing PDF text was incorporated when available, and all PDFs were processed using the same pipeline. We consider this a practical system-level comparison between text-based and page-image retrieval, rather than an evaluation of PDF extraction methods.

Fine-Tuning Details. We fine-tune the retriever using the ColBERT late-interaction contrastive loss defined in Equation (2) with temperature $\tau = 0.02$ and LoRA adapters (r=32, alpha=32, dropout = 0.1). In the text-component-only setting, LoRA targets the query encoder’s attention projections (q_proj, k_proj, v_proj, and o_proj), feed-forward projections (down_proj, gate_proj, and up_proj), and custom_text_proj; the vision encoder remains frozen. In the text+visual setting, the same attention and feed-forward projections are adapted in both the textual and visual transformer blocks, where the visual component denotes the blocks encoding page-image patches. All untargeted parameters remain frozen. Training is performed for 3 epochs with a learning rate of 5×10^{-5} and 100 warmup steps, using a per-device batch size of 16 on 4 NVIDIA L40S GPUs and no gradient accumulation, resulting in an effective batch size of 64. All runs use bfloat16 precision. Each fine-tuning configuration is repeated with three seeds (42, 43, and 44), and results are reported as mean $\pm$ standard deviation.

Evaluation Setup and Metrics. We evaluate retrieval at the page level: for each query, the retriever ranks all document pages in the benchmark against the gold relevant page. Results are reported separately for English, French, German, and Luxembourgish queries. Our main metric is nDCG@10; for analyses across retrieval depths, we additionally report nDCG@ k for $k \in \{1, 3, 5, 10\}$. Indexing time per page and query latency per query are aggregated across the four query languages and reported as mean $\pm$ standard deviation over three runs. All evaluation runs, including effectiveness and efficiency measurements, use one NVIDIA L40S GPU and eight CPU cores.

Retriever	Query Language				Avg.	Index (ms/page) ↓	Latency (ms/query) ↓
	EN	FR	DE	LB			
TEXT-ONLY RETRIEVERS							
bge-m3	72.48	71.63	72.47	76.12	73.18	21.714 ± 0.015	3.859 ± 0.027
Qwen3-Embedding-0.6B	73.02	71.63	75.21	67.97	71.96	42.863 ± 0.115	6.935 ± 0.008
multilingual-e5-large	75.96	75.02	77.24	68.32	74.14	11.068 ± 0.070	3.858 ± 0.012
jina-embeddings-v4	76.51	75.39	77.30	73.86	75.76	113.007 ± 0.287	26.271 ± 0.336
MULTI-VECTOR PAGE-IMAGE RETRIEVERS							
colSmol-256M	71.61	59.45	52.76	55.11	59.73	367.108 ± 1.415	3.914 ± 0.035
colSmol-500M	73.90	67.56	65.92	63.87	67.81	361.103 ± 11.169	4.125 ± 0.017
colpali-v1.3	77.03	77.43	78.46	75.78	77.18	61.801 ± 0.110	2.997 ± 0.007
colqwen2.5-v0.2	<u>78.07</u>	78.71	79.92	78.60	78.83	133.876 ± 0.813	4.892 ± 0.113
colnomic-embed-multimodal-3b	78.73	<u>78.44</u>	<u>79.59</u>	<u>77.89</u>	<u>78.66</u>	145.193 ± 0.277	4.981 ± 0.067

Table 2: Cross-lingual retrieval over Luxembourgish document pages. Query-language columns report nDCG@10 where best results are shown in **bold**, and second-best results are underlined. Efficiency values are mean ± standard deviation over three runs on the same hardware.

6 Results and Analysis

Our experiments evaluate cross-lingual retrieval over Luxembourgish document pages. We first compare text-only and page-image retrievers, then analyze query-language-specific fine-tuning, including an ablation on the visual component, and finally evaluate multilingual fine-tuning.

6.1 Page-Image vs. OCR-Based Text Retrieval over Luxembourgish Documents

We compare multilingual OCR-based text-only embedding baselines, which index page-level text produced by the OCR, layout-analysis, and table-extraction pipeline described in Section 5, with late-interaction page-image retrievers ranging from lightweight models below 500M parameters to models of up to 3B parameters. Table 2 shows that the strongest page-image retrievers outperform the evaluated text-only baselines across all query languages. Since the two families differ in input representation, extraction pipeline, and model architecture, we interpret this as a practical system-level comparison rather than a controlled modality ablation.

Performance remains strongly model-dependent. The smaller colSmol models underperform strong multilingual text-only retrievers such as jina-embeddings-v4. In contrast, colqwen2.5-v0.2 and colnomic-embed-multimodal-3b outperform all evaluated text-only models across every query language, while colpali-v1.3 also performs strongly on English, French, and German. This shows that sufficiently capable multi-vector page-image retrievers can outperform multilingual text em-

bedding models in cross-lingual retrieval. The best page-image retriever achieves an average nDCG@10 of 78.83, improving by 3.07 points over the strongest text-only baseline. A more customized OCR and layout-detection pipeline could nevertheless affect this gap.

These results are notable because the text-only baselines are explicitly designed for multilingual retrieval, whereas the page-image retrievers primarily target visual-document retrieval. Their strong zero-shot cross-lingual performance may partly reflect the multilingual VLM backbones used by several models.

Table 4 further examines performance by QA source. The page-image retrievers outperform the strongest text-only baseline on both the text-focused and visually grounded subsets, with larger gains on the visually grounded questions. This provides additional evidence that page-image retrieval is beneficial beyond the predominantly text-focused portion of the benchmark. However, the visually grounded subset contains only 49 QA pairs and yields higher scores for all retrievers, so these results remain descriptive and should be interpreted cautiously.

Efficiency. Table 2 shows clear efficiency differences across retrievers. Among text-only models, multilingual-e5-large indexes fastest, while jina-embeddings-v4 is slower, likely due to its larger multimodal backbone. Among page-image retrievers, colpali-v1.3 is the fastest, possibly because of its fixed-resolution processing and fixed visual-token budget. The slower ColSmol indexing further shows that runtime depends on visual-token count

Training setup	Query language				
	EN	FR	DE	LB	Avg.
Baseline	78.73	78.44	79.59	77.89	78.66
FT (EN)	81.66 ± 0.33 (+2.93)	81.34 ± 0.40 (+2.89)	81.84 ± 0.11 (+2.24)	78.48 ± 0.65 (+0.60)	80.83 ± 0.23 (+2.17)
FT (FR)	81.11 ± 0.57 (+2.38)	81.50 ± 0.49 (+3.06)	81.76 ± 0.40 (+2.17)	79.45 ± 0.75 (+1.57)	80.96 ± 0.55 (+2.29)
FT (DE)	79.48 ± 0.27 (+0.75)	80.51 ± 0.88 (+2.07)	81.56 ± 0.34 (+1.97)	79.34 ± 0.33 (+1.45)	80.22 ± 0.14 (+1.56)
Avg. FT	80.75 ± 0.15	81.12 ± 0.22	81.72 ± 0.08	79.09 ± 0.34	80.67 ± 0.11

Table 3: nDCG@10 by query language (columns) for the baseline model and models fine-tuned on different training languages (rows). Values are reported as mean $\pm$ standard deviation over three runs; values in parentheses indicate gains over the baseline on the same query language. Best values are shown in **bold**.

Retriever	Text-focused	Visually grounded	Overall
jina-embeddings-v4	75.27	81.10	75.76
colqwen2.5-v0.2	78.42	83.23	78.83
colnomic-embed-multimodal-3b	77.91	86.82	78.66

Table 4: Descriptive nDCG@10 split by QA source, averaged across query languages.

and implementation choices, not only parameter count. Query latency remains similar across most page-image models because page representations are precomputed.

For the text-only retrievers, page parsing introduces an additional shared preprocessing cost. Using the extraction pipeline described in Section 5, parsing takes 5.382 ± 0.054 s/page across three runs. Including this cost, the end-to-end preprocessing and indexing time ranges from 5.393 ± 0.054 s/page for multilingual-e5-large to 5.495 ± 0.054 s/page for jina-embeddings-v4. Thus, for text-only retrieval, document parsing dominates the overall offline preprocessing cost.

6.2 Cross-Lingual Transfer of Query-Language-Specific Fine-Tuning

This subsection studies how fine-tuning transfers across evaluation languages for the late-interaction page-image retrievers considered in this work. We fine-tune the retriever separately on English, French, and German queries, and evaluate each variant on all four query languages. We use colnomic-embed-multimodal-3b for this analysis, as both the baseline model and the backbone for fine-tuning, because its performance remains very close to the top model in our experiments, while ViDoRe leaderboard (Loison et al., 2026) results suggest it is a strong page-image retriever overall.

Table 3 shows that all language-specific fine-tuning settings improve over the baseline on every query language. FT (FR) achieves the highest over-

all average at 80.96, closely followed by FT (EN) at 80.83, while FT (DE) reaches 80.22. The largest language-specific gains are obtained by FT (EN) on English (+2.93) and FT (FR) on French (+3.06). Cross-lingual transfer is also substantial: FT (EN) achieves the highest German score at 81.84 (+2.24), while FT (FR) performs best on Luxembourgish at 79.45 (+1.57).

These results show that the strongest configuration does not always correspond to matching the fine-tuning and evaluation languages. In particular, English fine-tuning transfers strongly to German, while French fine-tuning provides the strongest Luxembourgish performance. German fine-tuning improves all four languages but produces the lowest overall average among the three variants. Overall, query-language-specific fine-tuning consistently improves multilingual retrieval, while the magnitude of transfer depends on the training and evaluation language pair.

Figure 2 shows that these trends are broadly maintained across retrieval cutoffs. FT (EN) remains strongest for English queries, whereas the fine-tuned variants are much closer for French and German. For Luxembourgish, FT (FR) and FT (DE) generally perform better than FT (EN). The uncertainty bands also overlap in several cases, indicating that small differences between fine-tuning languages should not be over-interpreted. Overall, the improvements persist across different values of k , showing that the cross-lingual benefits of fine-tuning are not limited to nDCG@10.

Effect of Updating the Visual Component. In all previous fine-tuning experiments, we use a setting in which both the textual and visual components of the page-image retriever are updated. To isolate the contribution of visual adaptation, we compare this setting with the corresponding text-

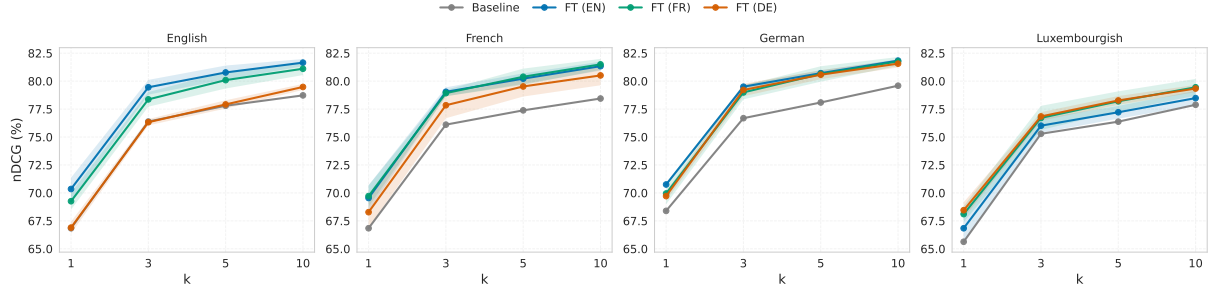


Figure 2: Multilingual retrieval performance (nDCG@ k) across query languages. Shaded bands indicate ± 1 standard deviation across three fine-tuning runs.

Training setup	Query language			
	EN	FR	DE	LB
FT (EN)	1.69 $\pm$ 0.57	1.49 $\pm$ 0.53	1.41 $\pm$ 0.12	0.21 $\pm$ 0.38
FT (FR)	2.27 $\pm$ 0.16	1.32 $\pm$ 0.34	1.78 $\pm$ 0.44	1.27 $\pm$ 0.98
FT (DE)	2.49 $\pm$ 1.25	2.53 $\pm$ 1.46	1.34 $\pm$ 1.07	0.82 $\pm$ 0.48

Table 5: Difference in nDCG@10 between text+visual and text-component-only fine-tuning across query languages. Values are reported as mean $\pm$ standard deviation over three runs; positive values indicate gains from updating the visual component. Best values are shown in **bold**.

component-only fine-tuning variant. Table 5 reports the paired difference in nDCG@10 between the two settings across three runs.

The mean differences are positive for every training and evaluation language combination, indicating that updating the visual component provides an additional benefit beyond text-component-only adaptation. The magnitude of this gain varies substantially across settings, ranging from 0.21 points for FT (EN) evaluated on Luxembourgish to 2.53 points for FT (DE) evaluated on French. The larger gains observed for some FT (DE) configurations are accompanied by higher run-to-run variability, while several EN and FR configurations show smaller but more stable improvements. Overall, these results suggest that visual adaptation provides complementary gains, although its contribution depends on both the fine-tuning language and the evaluation language.

6.3 Multilingual Fine-Tuning for Improved Cross-Lingual Retrieval

We further investigate whether multilingual fine-tuning improves cross-lingual retrieval beyond the single-language fine-tuning studied earlier. We compare the best single-language setting, FT (FR), with two multilingual variants: one fine-tuned on English, French, and German, and an-

other fine-tuned on English, French, German, and Luxembourgish. We again use colnomic-embed-multimodal-3b as both the baseline model and the backbone used for fine-tuning.

As shown in Table 6, multilingual fine-tuning on English, French, and German underperforms FT (FR) across all query languages, with an average decrease of 1.03 nDCG points. The largest decrease occurs on Luxembourgish queries (-1.97). This degradation may partly reflect the smaller number of training examples available per language in the multilingual setting compared with single-language FT (FR), which can reduce language-specific adaptation. The absence of Luxembourgish supervision may further contribute to the larger drop observed on Luxembourgish queries.

Adding Luxembourgish substantially changes this pattern. The EN+FR+DE+LB setting reaches an average nDCG@10 of 80.94, essentially matching FT (FR) at 80.96. Relative to EN+FR+DE, adding Luxembourgish improves performance across all query languages, with the largest increase on Luxembourgish queries, from 77.48 to 80.02 ($+2.54$ points). Compared with FT (FR), the multilingual setting remains close on English, French, and German, with differences of -0.08 , -0.46 , and -0.10 , respectively, while improving Luxembourgish retrieval by $+0.57$.

Since the corpus pages are in Luxembourgish, direct exposure to Luxembourgish queries may help align query representations with the corresponding document pages. However, the multilingual settings also differ from single-language fine-tuning in the amount and distribution of training data per language, so the effect cannot be attributed to Luxembourgish supervision alone. Overall, the results show that multilingual fine-tuning including Luxembourgish can recover the performance lost in the EN+FR+DE setting, match the strongest

Training setup	Query language				
	EN	FR	DE	LB	Avg.
Baseline	78.73	78.44	79.59	77.89	78.66
FT (FR)	81.11 ± 0.57	81.50 ± 0.49	81.76 ± 0.40	<u>79.45 ± 0.75</u>	80.96 ± 0.55
Multilingual FT					
EN+FR+DE	80.47 ± 0.39 (-0.63)	80.46 ± 0.16 (-1.04)	81.30 ± 0.26 (-0.46)	77.48 ± 0.48 (-1.97)	79.93 ± 0.12 (-1.03)
EN+FR+DE+LB	<u>81.03 ± 0.51</u> (-0.08)	<u>81.04 ± 0.44</u> (-0.46)	<u>81.66 ± 0.44</u> (-0.10)	80.02 ± 0.56 (+0.57)	<u>80.94 ± 0.41</u> (-0.02)

Table 6: nDCG@10 across query languages for the baseline, the best single-language fine-tuning setting, and multilingual fine-tuning variants. Results are reported as mean $\pm$ standard deviation over three runs. Values in parentheses indicate changes relative to FT (FR); the best value is shown in **bold** and the second-best is underlined.

single-language configuration overall, and provide the strongest Luxembourgish retrieval performance. Given the variability across three runs, the small differences between FT (FR) and EN+FR+DE+LB should be interpreted as comparable overall performance rather than a clear advantage for either setting.

7 Conclusion

Luxembourgish provides a representative low-resource use case for studying cross-lingual retrieval, given the scarcity of dedicated resources and retrieval benchmarks for the language. In this work, we introduced *LëtZCross*, a cross-lingual benchmark for page-level retrieval over Luxembourgish PDF documents in a page-image RAG setting. The benchmark combines automatically generated text-focused QA pairs with manually authored visually grounded QA pairs. Our results show that late-interaction page-image retrievers outperform the evaluated OCR-based text-only baselines in a system-level comparison, and that fine-tuning improves retrieval with gains that transfer across query languages. Among the evaluated single-language settings, French fine-tuning yields the highest mean performance on Luxembourgish queries, while among the multilingual configurations, including Luxembourgish produces the strongest results and substantially improves Luxembourgish-query retrieval. Together, these findings provide a first benchmark and empirical study for cross-lingual page-level retrieval over Luxembourgish document pages.

Limitations

This work has several limitations. First, *LëtZCross* is relatively small, with a limited manually authored visually grounded subset compared with the automatically generated text-focused QA pairs.

Consequently, the evaluation is dominated by text-focused queries, and results on the visually grounded subset should be considered complementary rather than representative of the benchmark. Second, the comparison between text-only and page-image retrievers is system-level: the text-only baselines rely on document parsing with OCR, layout analysis, and table extraction, while page-image retrievers operate on rendered pages. Alternative parsing or hybrid pipelines could therefore affect the observed gap. Third, all non-English queries are translated from English rather than independently authored, which may introduce translation artifacts or English-oriented formulations. Moreover, query translation and automatic QA generation rely on LLMs, although we apply validation and manual review to improve consistency and quality. Finally, our findings are specific to Luxembourgish and should be validated on other low-resource languages before broader generalization.

Acknowledgments

This research was funded in whole or in part by the Luxembourg National Research Fund (FNR), grant reference NCER22/IS/16570468/NCER-FT. We acknowledge Google.org for providing Gemini credits used in this work. We also thank the anonymous reviewers for their careful reading and constructive feedback.

References

- Voyage AI. 2026. Voyage-4-nano. <https://huggingface.co/voyageai/voyage-4-nano>. State-of-the-art text embedding model with 32k token context length.
- Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master’s thesis, University of Trier.

- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.
- Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.
- Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023. Lux-asr: Building an asr system for the luxembourgish language. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.
- Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–596.
- Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. 2025. [jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval](#).
- Chao-Wei Huang, Chen-An Li, Tsu-Yuan Hsu, Chen-Yu Hsu, and Yun-Nung Chen. 2024. Unsupervised multilingual dense retrieval via generative pseudo labeling. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 736–746.
- Vladimir Karpukhin, Barlas Oguz, 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.
- Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. Finepdfs. <https://huggingface.co/datasets/HuggingFaceFW/finepdfs>.
- António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. 2026. [Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios](#).
- Cedric Lothritz, Bertrand Lebigot, 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 LREC*.
- Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhui Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.
- Nomic AI. 2025. Colnomic embed multimodal 3b. <https://huggingface.co/nomic-ai/colnomic-embed-multimodal-3b>. Hugging Face model card.
- Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. [Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish](#). In *Proceedings of SIGUL (LREC-COLING)*.
- Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. [LuxBank: The First Universal Dependency Treebank for Luxembourgish](#). In *Proceedings of TLT*.
- Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. [Text Generation Models for Luxembourgish with Limited Data: A Balanced Multilingual Strategy](#). In *Proceedings of VarDial (COLING)*.
- Christoph Purschke. 2020. [Attitudes toward multilingualism in luxembourg. a comparative analysis of online news comments and crowdsourced questionnaire data](#). *Frontiers in Artificial Intelligence*, 3.
- Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. [Publish or hold? automatic comment moderation in Luxembourgish news articles](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 968–978, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL (LREC)*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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*.