# Lost in Retrieval

How much does RAG degrade across languages, and which fix closes the gap?

Lisann Jill Morlok
M.Sc. Informatics, Technical University of Munich

November 2026

**Abstract.** Retrieval-augmented generation (RAG) systems are usually built and tested in English, while many users ask in other languages over mostly English knowledge bases. I measure this gap under controlled conditions on XQuAD, where every German and Spanish question has an exact gold English paragraph. A naive English retriever ranks the gold paragraph first for 88.5% of English questions but only 36.9% of German and 34.0% of Spanish ones. Query translation, multilingual embeddings and cross-encoder reranking each recover most of the gap, and they stack: multilingual embeddings with reranking reach 95.8% (German) and 97.6% (Spanish). An error analysis shows that the naive systems succeed almost only through words shared verbatim with the English paragraph. It also explains a surprising result: BM25 does worse on Spanish than on German because Spanish articles such as *el* and *los*, rare in English text, receive high weights and pull queries towards the wrong paragraphs. At answer level, correctness follows retrieval, and the generator abstains rather than inventing an answer in 43 of 44 retrieval failures. An LLM judge agrees with an independent model's labels on correctness ($\kappa = 0.98$), but its faithfulness scores could not be validated.

## 1 Introduction

A RAG system answers a question in two steps: it retrieves passages from a document collection, then asks a language model to answer from them [3]. If the retriever cannot find the right passage, the generator cannot recover. Cross-lingual use puts the most strain on this first step. A German question over an English corpus shares few surface words with the passage that answers it, which is what both keyword search and English-only embedding models depend on.

There are three common fixes: translate the query into the corpus language, embed queries and passages in a shared multilingual space, or rerank candidates with a multilingual cross-encoder. This study compares them on the same questions, asks whether their differences are statistically real, follows them through to generated answers, and analyses which questions still fail and why.

**Contributions.** (1) A controlled comparison of six retrieval conditions on 1,190 questions in each of three languages, with bootstrap confidence intervals and paired significance tests. (2) An answer-level evaluation with an LLM judge that is validated, per language, before its scores are used. (3) An error analysis that tests a stated hypothesis, rejects it, and identifies the actual mechanism with a direct intervention. All code, configurations and per-query results are in the accompanying repository.

## 2 Setup

**Benchmark.** XQuAD [1] is a professionally translated subset of SQuAD v1.1: 240 English paragraphs and 1,190 questions, each available in German and Spanish. Question ids are shared across languages, so each German or Spanish question maps to an exact gold English paragraph. The corpus is the 240 English paragraphs. English queries give the same-language upper bound.

**Retrieval conditions.** All conditions share one interface.

1. **BM25** [5], the lexical baseline.
2. **English-only embeddings**: `all-MiniLM-L6-v2` [4, 7], a typical English RAG stack used naively.

| Condition | English | German | Spanish |
|---|---|---|---|
| BM25 | 0.918 [0.902, 0.934] | 0.371 [0.341, 0.400] | 0.192 [0.171, 0.213] |
| English-only embeddings | 0.885 [0.866, 0.903] | 0.369 [0.340, 0.396] | 0.340 [0.314, 0.366] |
| Translate, then English emb. | — | 0.854 [0.834, 0.873] | 0.846 [0.826, 0.866] |
| Multilingual `me5-base` | 0.955 [0.943, 0.967] | 0.903 [0.885, 0.919] | 0.917 [0.900, 0.933] |
| Multilingual `bge-m3` | 0.940 [0.926, 0.954] | 0.904 [0.887, 0.921] | 0.918 [0.901, 0.934] |
| `me5-base` + reranker | **0.987** [0.980, 0.992] | **0.958** [0.946, 0.968] | **0.976** [0.966, 0.984] |

**Table 1:** Recall@1 (gold paragraph ranked first), 1,190 questions per language, 95% bootstrap intervals. Recall@5/10 and MRR are in the repository.

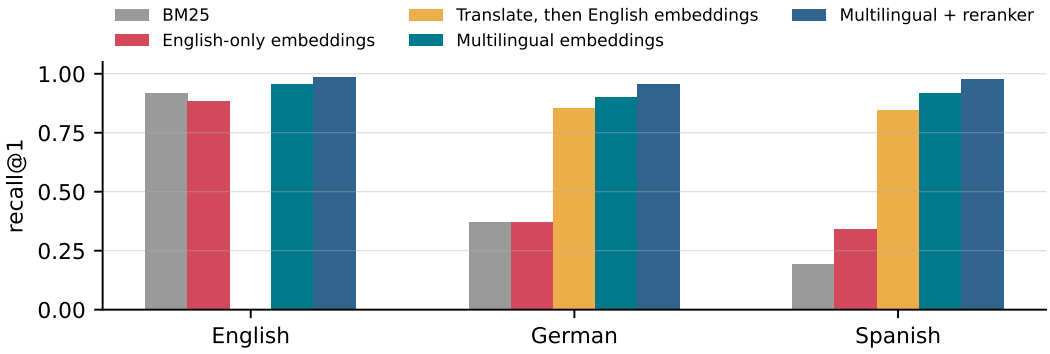



**Figure 1:** Recall@1 by condition and query language.

3. **Translate, then retrieve**: OPUS-MT [6] translates the query into English, followed by condition 2.
4. **Multilingual embeddings**: `multilingual-e5-base` (278M parameters) [8] and `bge-m3` (568M) [2].
5. **Multilingual embeddings + reranker**: `bge-reranker-v2-m3` reorders the top 20 candidates from `multilingual-e5-base`.

**Metrics.** Recall@$k$ ($k \in \{1, 5, 10\}$) and MRR, with 95% percentile bootstrap intervals over questions (2,000 resamples). Two conditions are compared with a two-sided paired bootstrap test on per-question differences (10,000 resamples).

**Answers.** For the naive condition (English-only embeddings) and the best one (multilingual embeddings + reranker), a local LLM (`qwen3.8:27b-mlx`, temperature 0) answers 50 randomly sampled questions per language from the top three paragraphs. The same 50 ids are used for every language and condition. The prompt asks for a short answer, or exactly "I don't know" if the context does not contain one. The same model, as an LLM judge [9], scores each answer for *correctness* (matches a reference answer) and *faithfulness* (every claim is supported by the context).

## 3 Results

### 3.1 Retrieval

Table 1 and Figure 1 show the main result. Used naively, the English-only model loses more than half its accuracy across languages (0.885 → 0.369 for German, 0.340 for Spanish). Every fix recovers most of this gap. Table 2 shows that each step is significant except one: `bge-m3`, twice the size of `multilingual-e5-base`, is statistically indistinguishable from it ($p > 0.86$). Reranking adds a further 5–6 points and brings both languages within 1–3 points of English.

| Comparison (recall@1) | German | Spanish |
|---|---|---|
| Translate vs. naive English embeddings | $+0.485$, $p < 10^{-4}$ | $+0.506$, $p < 10^{-4}$ |
| Multilingual `me5-base` vs. translate | $+0.049$, $p < 10^{-4}$ | $+0.071$, $p < 10^{-4}$ |
| `bge-m3` vs. `me5-base` | $+0.002$, $p = 0.86$ | $+0.001$, $p = 0.91$ |
| Reranker vs. `me5-base` alone | $+0.055$, $p < 10^{-4}$ | $+0.059$, $p < 10^{-4}$ |

**Table 2:** Paired bootstrap tests on the same questions.

| Condition | Lang | Gold paragraph in top 3 | Correct (judge) |
|---|---|---|---|
| Naive (English-only embeddings) | en | 0.98 | 0.96 [0.90, 1.00] |
| | de | 0.60 | 0.58 [0.44, 0.72] |
| | es | 0.54 | 0.50 [0.36, 0.64] |
| Best (`me5-base` + reranker) | en | 1.00 | 0.98 [0.94, 1.00] |
| | de | 1.00 | 0.98 [0.94, 1.00] |
| | es | 0.98 | 0.92 [0.84, 0.98] |

**Table 3:** Answer correctness on 50 questions per language (same ids throughout).

Because first-stage recall@10 is already $\geq 0.994$ cross-lingually, the reranker's 20-candidate cap is not a limit.

### 3.2 Answers

Answer correctness closely tracks whether the gold paragraph was retrieved (Table 3). In the naive setup, 44 of the 150 answers had no gold paragraph in their context. For 43 of them the model replied "I don't know". Once it answered anyway: asked about Fort Duquesne, it answered "San Mateo", a name taken from a paragraph about a different fort. In this setup, the cross-lingual gap is therefore decided at retrieval and mostly appears as abstention, not hallucination.

### 3.3 Can the judge be trusted?

Before the judge's scores were used, 150 answers (50 per language, both conditions mixed and shuffled, condition hidden) were labelled against written guidelines, blind to the judge. The labels were produced by Claude (Anthropic), a model of a different family from the judge. This measures cross-model agreement, which is a weaker check than human labels, and it is reported as such.

For correctness, agreement is 149/150 ($\kappa = 0.977$ overall; 1.00 for German and English, 0.95 for Spanish). The one disagreement is a consistency slip, not a language-specific failure: the judge accepted "water, coal and hydrocarbons" in German and English but not in Spanish. For faithfulness, the sample contains a single unfaithful answer (the "San Mateo" case), and the judge marked it faithful. With one negative, $\kappa$ is 0 and uninformative. The judge's faithfulness scores (100% faithful) are therefore not used as evidence.

## 4 Error analysis

Features are computed from each question's English original, so they are identical across languages. An *anchor token* is a query word of at least four characters, or any token containing a digit, which can match an English paragraph by surface form alone. A query's *overlap* is the share of its anchor tokens that occur verbatim in the gold paragraph.

**Naive retrievers work only through shared words.** Half of the German and Spanish questions (51%) have no anchor token in common with their gold paragraph, against 1% for English. On those questions, BM25 finds the gold paragraph 1–2% of the time and English-only

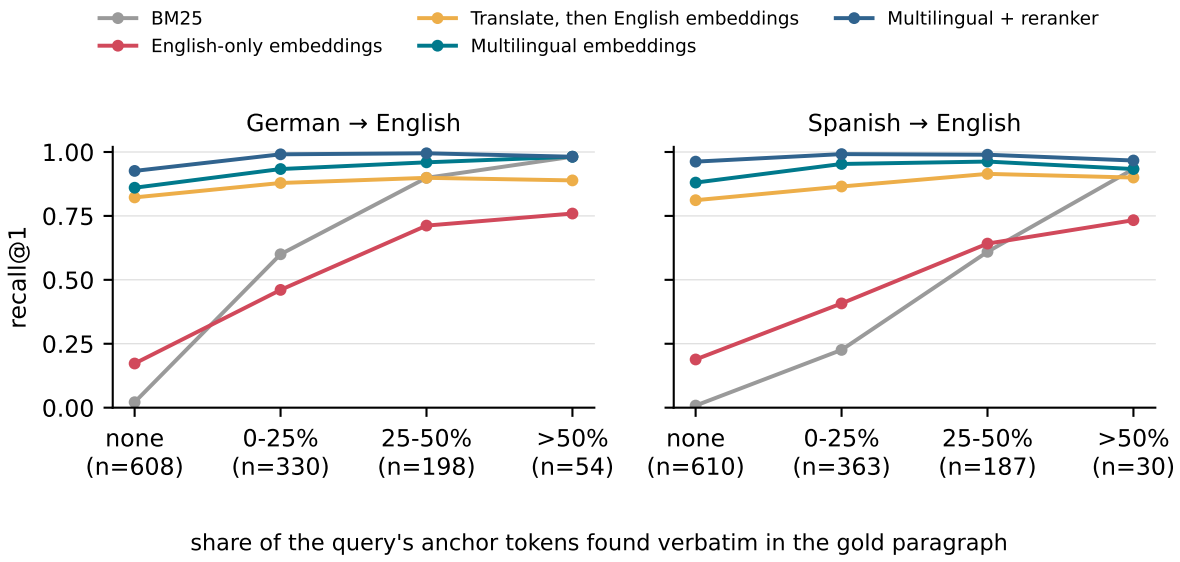



**Figure 2:** Recall@1 by lexical overlap between the (untranslated) query and its gold paragraph. Naive retrievers depend on overlap; multilingual ones barely do.

embeddings 17–19% (Figure 2). Both rise steeply with overlap. Translation and the multilingual retrievers are almost flat: with reranking, recall@1 is 0.93 (German) and 0.96 (Spanish) even with no shared words.

**Spanish function words, not German loanwords, explain the BM25 gap.** BM25 is significantly better on German than Spanish (0.371 vs. 0.192; paired $p < 10^{-4}$). My initial hypothesis was that German questions keep more English-identical words. The data rejects it: mean overlap is almost the same (0.14 vs. 0.13), and so is the share of questions with none (51% for both). The difference lies in what drives a miss. For each BM25 top-1 miss, I found the highest-IDF word that the query shares with the wrongly ranked paragraph. In 93% of Spanish misses it is a short function word: *el* (318 misses), *la* (231), *los* (133), *de* (83). These words occur in only 3–10 of the 240 English paragraphs, in names like *Los Angeles* or *El Niño*, so BM25 weights them heavily and sends many Spanish queries to the same few paragraphs. German misses are more often caused by having no shared word at all (22%). Dropping short words from the queries is a direct test, and it confirms the mechanism. Spanish BM25 recall@1 nearly doubles (0.19 → 0.36), German is unchanged (0.37 → 0.37), and the gap closes.

**Translation fails when it loses names.** Over questions whose English original contains capitalised terms, translations that led to a miss preserved 82–83% of those terms, against 92% for translations that led to a hit. Translation and multilingual embeddings also fail on different questions: 113 German and 125 Spanish questions are missed by translation but found by `me5-base`, and 55 and 41 the other way round. That suggests an ensemble as a cheap improvement.

**Question type matters less than overlap.** For the best system, "how" and "where" questions are hardest (0.89–0.91 in German), and questions whose answer is a number or date are easiest (0.98–0.99). For the naive system, no question or answer type comes close to closing the gap.

**Remaining errors are partly not cross-lingual.** The best system misses 50 German and 29 Spanish questions at rank 1. Of these, 22% and 48% are also missed in English, so they are hard questions rather than translation problems. In 12–17% of misses, the paragraph ranked first also contains the gold answer string, so not every miss would cost a correct answer. There are 54 misses that occur only cross-lingually. Several of the first ten are underspecified without their original paragraph ("When were *these* settlers naturalized?", "What were the court proceedings

about?"), and one was translated loosely ("how big" for "how busy" an airport is). The full list is in the repository.

## 5 Limitations

- The corpus is small (240 paragraphs), so absolute recall is high for any reasonable model. The informative quantity is the gap between conditions. A harder setting would add distractor paragraphs, e.g. from MLQA or Wikipedia.
- Translated benchmark questions are cleaner than real user queries, and all three languages are European and written in Latin script.
- Answer-level results use 50 questions per language, so their intervals are wide (about $\pm 0.14$ for the naive setup).
- The same model generates and judges answers, which can bias the judge towards its own outputs. The judge was validated against a second model, not a human, and its faithfulness scores are unvalidated.
- Anchor tokens and answer types are simple heuristics. They are adequate for the comparisons made here, but not a linguistic analysis.

## 6 Conclusion

Used naively, a typical English RAG stack loses about 60% of its retrieval accuracy on German and Spanish questions, because it can only match words that happen to survive translation. Multilingual embeddings close most of the gap, a larger embedding model does not help, and a multilingual reranker closes most of what is left. Because the generator abstains when the context lacks the answer, retrieval quality largely determines answer quality in this setup. The error analysis shows the value of checking a plausible explanation: the German–Spanish difference for BM25 comes not from shared vocabulary but from Spanish articles that the English corpus treats as rare and informative. Next steps are human labels for the judge, a faithfulness test set enriched for unsupported answers, a distractor corpus, and a translation–embedding ensemble.

## References


[1] M. Artetxe, S. Ruder, D. Yogatama. On the cross-lingual transferability of monolingual representations. *ACL*, 2020.

[2] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Findings of ACL*, 2024.

[3] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 2020.

[4] N. Reimers, I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP*, 2019.

[5] S. Robertson, H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009.

[6] J. Tiedemann, S. Thottingal. OPUS-MT: Building open translation services for the world. *EAMT*, 2020.

[7] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 2020.

[8] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei. Multilingual E5 text embeddings: A technical report. *arXiv:2402.05672*, 2024.

[9] L. Zheng et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets and Benchmarks*, 2023.