

GutenBench: A Taxonomy-Based Dataset for Evaluation of Information Retrieval Models

Márcio Monteiro¹, Denys Senkin¹, Abhilasha Ravichander²,
Marius Kloft¹, Sophie Fellenz¹

¹RPTU University of Kaiserslautern-Landau, ²Max Planck Institute for Software Systems

Correspondence: marcio.monteiro@cs.rptu.de

Abstract

While retrieval-augmented generation (RAG) relies on retrieved evidence to mitigate hallucination, its reliability is constrained by retrieval quality and diagnosing where retrieval fails remains difficult. Existing benchmarks typically report a single aggregate score over heterogeneous query sets and are often dominated by queries whose relevant passages can be identified through simple lexical overlap, concealing the failure modes that matter most in practice. We introduce GUTENBENCH, a fine-grained, zero-shot evaluation dataset built around a four-level retrieval taxonomy: lexical, semantic, explicit aggregation, and latent conceptual retrieval. It pairs queries with relevance judgments adjudicated by subject-matter experts, enabling precise and controlled comparisons of retrievers. We evaluate several off-the-shelf retrievers in an asymmetric, English-to-target cross-lingual setting across 272 Bible translations (141 languages) and find that performance is strong on lexical and semantic queries, but degrades when models must aggregate information across passages or recognize conceptually relevant passages with no surface overlap.

1 Introduction

Large language models (LLMs) have revolutionized natural language processing, demonstrating remarkable capabilities in generating fluent and contextually relevant text. However, their deployment in high-stakes applications is often hindered by the risk of hallucinations, that is, generating plausible but factually incorrect or ungrounded statements (Thirunavukarasu et al., 2023).

Retrieval-augmented generation (RAG) directly addresses the problem of hallucinations by grounding the LLM’s answers in externally retrieved text, rather than relying solely on parametric memory (Lewis et al., 2020). This has made RAG a widely adopted strategy for reducing hallucinations (Shuster et al., 2021; Huang et al., 2025).

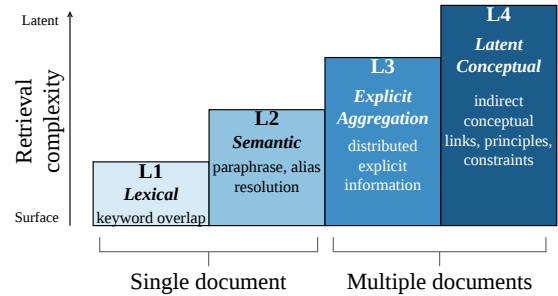


Figure 1: Taxonomy of retrieval capabilities in GUTENBENCH. Each step raises retrieval complexity and builds on the levels below: **L2** adds semantic matching to the lexical matching of **L1**, **L3** adds aggregating explicit information across multiple documents, and **L4** adds a latent conceptual link no passage states explicitly.

RAG’s ability to reduce hallucinations, however, strongly depends on the evidence it retrieves: without reliable retrieval, grounding fails, and with it RAG’s core promise. LLMs are easily distracted by noisy or semantically similar documents (Shi et al., 2023), and providing irrelevant or incomplete context makes hallucinations worse, not better (He et al., 2024; Zhang and Zhang, 2025; Huang et al., 2025; Joren et al., 2025).

Yet diagnosing retrieval failures requires more fine-grained evaluation: a system that performs well when results are averaged over heterogeneous queries may still fail on certain queries that require semantic matching, multi-document explicit aggregation, or latent conceptual retrieval. These failures are especially important for RAG, where missing or incomplete evidence can directly undermine the generated answer. Targeted retrieval evaluation, with sufficient resolution to distinguish these behaviors, is therefore essential for guiding model development and understanding the limits of RAG reliability.

Despite the proliferation of IR benchmarks, a critical diagnostic gap remains. Existing suites,

from predominantly English collections (Petroni et al., 2021; Thakur et al., 2021; Muennighoff et al., 2023) to comprehensive multilingual and cross-lingual suites (Zhang et al., 2023; Lawrie et al., 2025; Chen et al., 2025), report only aggregate metrics, making it difficult to attribute failures to specific retrieval behaviors. Moreover, the dominant construction practice (i.e., generating queries from sampled passages and pooling candidates with a lexical baseline) systematically biases benchmarks toward lexically transparent queries (Thakur et al., 2021), inflating the apparent competence of methods like BM25 (Robertson and Zaragoza, 2009) and hiding the conditions under which dense retrievers actually degrade. What is missing is a controlled, fine-grained benchmark that isolates distinct retrieval capabilities within a single unified corpus, enabling precise diagnosis of where and why retrievers fail.

To address this, we introduce GUTENBENCH¹, a fine-grained, zero-shot evaluation dataset for information retrieval.² As illustrated in Figure 1, GUTENBENCH categorizes queries into four distinct tiers of retrieval competencies, ranging from basic lexical lookups (L1) to latent conceptual retrieval (L4). It complements rather than replaces broad heterogeneous benchmarks, serving as a controlled diagnostic instrument that exposes failure modes aggregate-heavy evaluations obscure.

Our main contributions are:

- We introduce a four-level taxonomy that disentangles retrieval capabilities.
- We provide a compact benchmark with curated queries and relevance judgments adjudicated by subject-matter experts (SMEs), supporting controlled comparisons of retrievers in a zero-shot setting.
- We evaluate off-the-shelf retrievers across 272 Bible translations (141 unique languages), showing consistent degradation with retrieval complexity and substantial transfer gaps for low-resource languages.

Section 2 establishes the formal background and Section 3 defines the four-tier taxonomy. Sections 4

and 5 describe the benchmark construction pipeline and the resulting dataset. Section 6 presents the zero-shot evaluation of sparse and dense retrievers, Section 7 situates the work in the broader literature, and Section 8 concludes with future directions.

2 Background

Information retrieval systems are commonly evaluated as ranking models. Given a query Q and the retrieval corpus $\mathcal{D} = \{d_1, \dots, d_n\}$, retrievers must return a ranked list of relevant documents $\mathcal{D}_Q^+ = \{d_{Q,1}^+, \dots, d_{Q,m}^+\} \subset \mathcal{D}$, where $m \ll n$. Negative documents are defined as $\mathcal{D}_Q^- = \mathcal{D} \setminus \mathcal{D}_Q^+$.

The definition of relevance, however, is not fixed and is often based on human judgments. Query relevance judgments may be binary, assigning each query–document pair a label $\text{rel}_i \in \{0, 1\}$, or graded, assigning different levels of relevance, for example $\text{rel}_i \in \{0, 1, 2\}$.

To evaluate retrievers, we adopt normalized discounted cumulative gain (nDCG), one of the most widely used metrics for ranked retrieval (Järvelin and Kekäläinen, 2002; Thakur et al., 2021; Enevoldsen et al., 2025). It measures how well a retriever ranks highly relevant documents near the top of the candidate list while applying a logarithmic discount to lower-ranked items. For a ranked list of length k , the discounted cumulative gain is defined as

$$\text{DCG}@k = \sum_{i=1}^k \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)}, \quad (1)$$

where rel_i is the relevance score of the document at rank i . The normalized score is obtained by dividing by the ideal DCG, computed from the best possible ordering:

$$\text{nDCG}@k = \frac{\text{DCG}@k}{\text{IDCG}@k}. \quad (2)$$

Thus, $\text{nDCG}@k \in [0, 1]$, with higher values indicating that more relevant items are ranked closer to the top.

Query relevance judgments are often expensive to obtain because they require human assessment over query–document pairs, so judging every query in a large candidate pool is rarely feasible. Active learning reduces this burden by using a model to prioritize informative examples for annotation, such as uncertain, difficult, or potentially high-impact cases (Monarch, 2021).

¹GUTENBENCH is an abbreviation of Gutenberg Benchmark. The name is inspired by Johannes Gutenberg, who is considered the inventor of the modern printing press. The first major book printed with his invention was the Gutenberg Bible, which serves as the corpus used in this work.

²Dataset available at <https://huggingface.co/datasets/contemmc/gutenberg>.

3 A Taxonomy for Disentangling Retrieval Capabilities

Standard evaluation protocols report aggregate performance metrics over all queries, which can obscure systematic weaknesses: models with similar average scores may fail on entirely different query types, making it difficult to diagnose which retrieval behavior is responsible. GUTENBENCH addresses this by stratifying queries within a single controlled corpus along two dimensions: the number of supporting documents required, and the nature of the relevance link between query and relevant documents. In the following, we formalize this four-level taxonomy.

3.1 Single Evidence Sufficiency

Many queries can be fully supported by a single relevant document in the target corpus. Parallel documents (e.g., repeated definitions or accounts of the same event across different documents) may exist and should also be considered relevant, but any one of them is sufficient to support the query properly. How that document is found, however, depends on the type of link to the query. The two levels below capture this distinction.

L1 – Lexical. This level consists of queries whose relevant documents can be identified mainly through strong keyword or surface-form overlap between the query and the target text. For such cases, a lexical retriever is typically sufficient to identify relevant documents.

L2 – Semantic. At this level, keyword overlap alone is not sufficient to uncover relevant documents. Successful retrieval, instead, depends on semantic matching between the query and the target text, though the relevant document still directly supports the query’s information need. Typical examples include entity alias resolution, paraphrase recognition, synonymy, and category-member relationships.

3.2 Multi-Evidence Necessity

Not all queries, however, can be fully supported by a single relevant document from the target corpus. Sometimes, the information needed is distributed over different documents, and the several parts must be combined. Whether combining these parts is explicit or requires an inferred bridge separates the two levels below.

L3 – Explicit Aggregation. At this level, the query’s information need consists of several distinct facts or facets, each explicitly stated by some relevant document in the corpus. Because no single document states all of them, successful retrieval requires surfacing documents that together cover every part. Each one may itself be found through the lexical or semantic matching described before. What L3 adds, compared to L1 and L2, is the need to recognize a compound information need rather than rely on a single best match. Typical cases include relation composition, comparison, counting, event sequencing, and broad topical queries that are resolved mainly through synthesis of explicitly stated evidence.

L4 – Latent Conceptual Bridging. This level is the most complex and contains queries whose required evidence cannot be fully recovered through lexical overlap, ordinary semantic matching, or the aggregation of explicit claims alone. Successful retrieval requires an additional latent conceptual inference to establish why a document is relevant to the query: a governing principle, contextual constraint, implicit exception, analogy, symbolic connection, or a distinction that qualifies or reconciles conflicting passages. Unlike L3, where every required document contributes explicit information and ordinary aggregation is sufficient, L4 requires a conceptual bridge between the query and a document. Typical cases include interpretive framing, applying general principles to specific cases, and recognizing evidence relevant to apologetic or controversial disputes.

Table 1 illustrates each level with practical examples, explaining the relevance link between queries and their supporting passages. The proposed taxonomy organizes queries according to the retrieval capabilities they demand. This enables fine-grained diagnosis, unlike the aggregate reporting over heterogeneous queries that many benchmarks rely on. Hence, this also helps prevent highly unbalanced benchmarks, such as those dominated by L1 queries, from overstating the effectiveness of lexical retrievers, which may still fail on higher-level queries.

In the next section (Sec. 4), we describe how this taxonomy is operationalized in the construction of GUTENBENCH.

Example Query	Relevant Passages	Retrieval Rationale
L1: How many years were the children of Israel in Egypt?	<i>“Now the sojourning of the children of Israel, who dwelt in Egypt, was four hundred and thirty years.”</i> (Exodus 12:40)	High keyword overlap. The answer is explicitly stated in a single phrase.
L2: Who took Judas’s place among the twelve?	<i>“And they gave forth their lots; and the lot fell upon Matthias; and he was numbered with the eleven apostles.”</i> (Acts 1:26)	Single-document, but requires matching “taking a place” to being “numbered with” and “the twelve” to “apostles”.
L3: How many kings did Daniel serve during his time in Babylon?	A: <i>“This dream I king Nebuchadnezzar have seen. Now thou, O Belshazzar, declare the interpretation thereof...”</i> (Daniel 4:18) B: <i>“Then commanded Belshazzar, and they clothed Daniel with scarlet...”</i> (Daniel 5:29) C: <i>“So this Daniel prospered in the reign of Darius, and in the reign of Cyrus the Persian.”</i> (Daniel 6:28)	Requires retrieving a set of complementary documents that individually provide partial evidence and collectively satisfy the query’s information need. Note it also relies on L2 semantic alias resolution (Belshazzar is the Babylonian name of Daniel).
L4: Does the Bible portray <u>wealth</u> primarily as a sign of <u>divine favor</u> or a spiritual danger?	A: <i>“The blessing of the LORD, it maketh rich...”</i> (Proverbs 10:22) B: <i>“It is easier for a camel to go through the eye of a needle, than for a rich man to enter into the kingdom of God.”</i> (Matthew 19:24)	Evidence cannot be retrieved via explicit claims. The retriever must bridge latent inferential links between the query’s abstract concepts and passages that implicitly embody those principles.

Table 1: Qualitative examples of the four retrieval levels in GUTENBENCH, illustrating the distinct evidence types and retrieval capabilities required for each level.

4 Dataset Construction

We construct GUTENBENCH through a pipeline carefully designed to promote the reliability and completeness of its relevance judgments while keeping annotation costs manageable. Each query is triaged, annotated for relevance, and assigned the taxonomy level that its supporting evidence demands. This section describes each stage of this pipeline, which Figure 2 outlines.

4.1 Phase I: Query Collection, Sampling and Triage

The first step is to collect a large pool of candidate queries. Instead of asking annotators to reverse-engineer queries from sampled passages, a common protocol in reading comprehension datasets such as SQuAD (Rajpurkar et al., 2016), we start from questions collected from publicly available web content (refer to Appendix B for details). When queries are written with a known relevant passage in view, they may inherit the vocabulary, phrasing, or framing of that passage, creating document-conditioned query bias (Lee et al., 2019). They may also overrepresent single-document lookup tasks, since each query is often derived from one sampled passage.

Our approach mitigates this by drawing candidates from a heterogeneous set of external sources, reducing reliance on any single query-generation protocol. Although some source questions may still have been written with a specific passage in mind, the resulting candidate pool is broader, containing

questions that resemble what users naturally ask as well as questions that challenge retrieval models in different ways.

Sampling. Since the initial query pool is too large to annotate exhaustively,³ we use a sampling strategy inspired by selection criteria from active learning to decide which candidates advance to triage. This selection strategy aims to improve the diagnostic value of the annotated set, but it can also introduce sampling bias because the selected examples no longer follow the distribution of the original pool (Farquhar et al., 2021). We therefore combine two targeted sampling strategies with a random component, balancing diagnostic coverage with distributional coverage:

- **Diversity sampling.** To cover the query space broadly, we encode all candidate queries with a text embedding model, project the embeddings to a lower-dimensional space with UMAP, cluster them with HDBSCAN, and sample evenly across the resulting clusters. This prevents the most popular topics from dominating the selection, hence promoting diversity.
- **Disagreement sampling.** To identify queries that distinguish retrieval methods, we run a committee of dense retrievers over the entire pool and measure how much their top-ranked

³At the time of writing, the pool contains nearly 100K candidate queries.

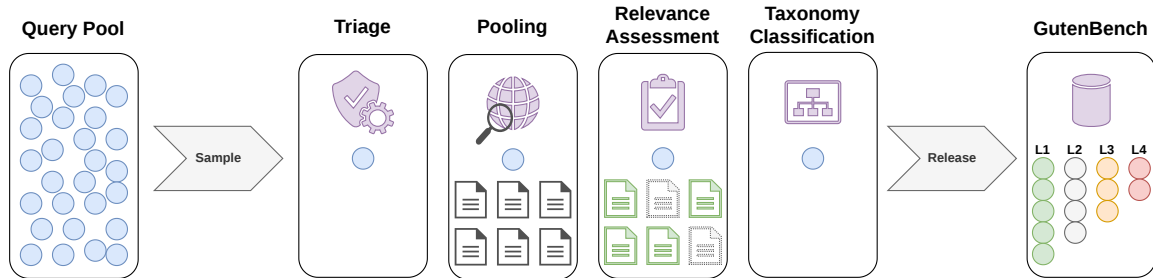


Figure 2: Overview of the GUTENBENCH construction pipeline. Sampled queries are screened at triage. For each accepted query, candidate passages are pooled and judged for relevance. Finally, the query is assigned a taxonomy level and released into the stratified benchmark. LLMs assist at every step; a human makes every final decision.

results overlap. Queries with low overlap are prioritized, as they are most likely to expose systematic differences between retrieval approaches.

- **Random sampling.** We include a random sample from the full candidate pool to keep the benchmark anchored to the pool’s overall distribution, preventing it from becoming overly adversarial or skewed toward edge cases.

Triage. After sampling, each candidate query undergoes AI-assisted triage before relevance assessment. The triage process serves as a first quality-control step: an AI agent provides a structured recommendation, but the dataset maintainers (the authors) make the final accept/reject decision. This stage checks whether the query is in domain, supported by the corpus, and suitable for retrieval evaluation, rejecting candidates that are unclear, underspecified, abusive, or focused on metadata rather than corpus content (e.g., *How many chapters are there in Book X?*). Overall, triage reduces annotation noise and improves consistency while preserving human control over benchmark inclusion.

4.2 Phase II: Candidate Passages Pooling and Relevance Assessment

This phase produces the relevance judgments for each accepted query. It proceeds in two steps: candidate passages are first gathered from multiple sources (pooling), and a human assessor then judges each one’s relevance to the query (assessment).

Candidate Passages Pooling. We construct each query’s judgment pool following the core idea of the TREC pooling methodology, in which candidate passages are collected from multiple systems

and later assessed by humans, rather than exhaustively judging the full corpus. Five frontier LLMs (ChatGPT, Claude, Gemini, Grok, and DeepSeek) propose candidate passages for each query, drawing on everything at their disposal, including parametric memory, tools, and web search. The pool further includes the top-5 results from a BM25 retriever, which add lexical candidates and nontrivial negatives, as well as any passages already proposed by the query’s original source when provided.

Relevance Assessment. A human assessor then grades every pooled passage for relevance to the query (0 for *not relevant*, 1 for *partially relevant*, and 2 for *definitely relevant*). To speed up this step, each candidate passage is presented together with the relevance votes of the different LLMs, which the assessor is free to follow or override (Faggioli et al., 2023). The human always has the final call and may also nominate relevant passages missing from the pool, extending the standard pooling protocol with manual pool expansion. Together, these mechanisms target broad coverage of each query’s relevant evidence.

4.3 Phase III: Document Mapping, Taxonomy Classification, and Dataset Release

In the final phase, each query is assigned one of the four taxonomy levels and incorporated into the benchmark.

Document Mapping. Once a query has its relevance judgments, the annotated passages are mapped into indexed documents: the corpus is segmented into fixed units delimited by the section headings of the New King James Version (NKJV) translation,⁴ and each passage is mapped to the doc-

⁴Editorial section headings divide the text into *pericopes*: short, self-contained thematic passages, each identified by a title and a verse range (e.g., *The History of Creation*, Genesis

ument that contains it. This aligns the final query relevance judgments (qrels) with the retrieval units indexed by the evaluated systems. The heading granularity offers a middle ground between two extremes: chunking by verse strips passages of their context, while chunking by chapter is too coarse and may exceed the context window of many retrieval models. Since several passages may fall within the same document, this mapping also produces the final set of relevant documents on which the level classification depends.

Level Classification. The GUTENBENCH maintainers then assign each query the lowest taxonomy level required to recover its supporting evidence, following the level definitions of Sec. 3, given the query and its relevant documents. As in Phase II (Sec. 4.2), the votes of several LLM judges are displayed as decision support: agreement among the judges corroborates clear-cut cases, while disagreement flags borderline queries for closer scrutiny. Nevertheless, a human always has the final call.

Release. Once its taxonomy level is decided, the query receives a topic identifier, following the TREC nomenclature, and becomes available in the next release of GUTENBENCH, with all underlying annotations preserved for future auditing.

Overall, multi-source pooling and assessor-driven expansion promote the completeness of the relevance judgments, human final decisions at every stage support their reliability, and LLM assistance keeps the annotation effort manageable. Moreover, because the target corpus is compact and stable, any judgment can be checked directly against the underlying text. Sec. 5 describes the resulting dataset.

5 Dataset Overview and Analysis

The current release of GUTENBENCH comprises 2,170 fully adjudicated queries, curated with the pipeline of Sec. 4, over the target corpus segmented into 2,813 non-overlapping chunks, which we call documents. Each query is assigned to one of the four taxonomy levels.

The distribution of these queries across the taxonomy is heavily skewed toward L1 (1,446 queries) and L2 (407 queries), while L3 (214 queries) and L4 (103 queries) remain comparatively rare. This

skew partly reflects the composition of the underlying query sources and the increasing difficulty of producing reliable judgments at higher levels.

Aggregated over this skewed distribution, scores would mostly reflect performance on L1 queries, allowing classical sparse baselines like BM25 to appear competitive with dense retrievers, as observed on existing benchmarks (Thakur et al., 2021; Sciavolino et al., 2021; Zhang et al., 2023). By reporting results per level, GUTENBENCH exposes failures beyond lexical matching (L2–L4) that such aggregates obscure. Each per-level score, however, is computed from a subset of the queries, smallest for L4. The bootstrap analysis in Appendix I, however, suggests that even this small subset yields stable diagnostic estimates, at least in the English setting.

6 Experiments and Results

To establish baselines and demonstrate GUTENBENCH’s diagnostic utility, we evaluate⁵ a suite of sparse and dense retrievers across all four taxonomy levels and 272 translations.

6.1 Experimental Setup

We evaluate models in a zero-shot, first-stage retrieval setting using nDCG@10 as the main metric. We structure the evaluation into two settings: an English retrieval setting, and an asymmetric English-to-target cross-lingual setting, with target languages grouped into three language resource tiers (high, mid, low) and scores macro-averaged over the translations in each tier. The English setting includes BM25 and a range of widely used dense retrievers, including top performers from the MMTEB leaderboard (Enevoldsen et al., 2025), whereas the cross-lingual setting includes only multilingual dense retrievers. In both settings, relevance judgments are shared across translations, since documents are delimited by the same verse ranges in every translation (Sec. 4.3). Crucially, all scores are stratified by taxonomy level to isolate specific capabilities, with $\pm$ denoting standard deviation across translations. For more information about the baselines, including implementation details, refer to Appendix J. All experiments in this section were performed on eight A100 40GB GPUs, totaling 367.2 GPU-hours.

1:1–2:7). Each pericope constitutes one document, yielding 2,813 documents in total.

⁵Evaluation code available at <https://github.com/contemmmcm/gutenbench>.

Model	Size	L1	L2	L3	L4
<i>Sparse model</i>					
BM25	–	0.712 $\pm$ 0.046	0.325 $\pm$ 0.026	0.306 $\pm$ 0.016	0.200 $\pm$ 0.015
<i>Open-weights models (<1B)</i>					
SBERT L6	20M	0.437 $\pm$ 0.027	0.206 $\pm$ 0.023	0.247 $\pm$ 0.013	0.216 $\pm$ 0.013
SBERT L12	30M	0.460 $\pm$ 0.022	0.231 $\pm$ 0.022	0.269 $\pm$ 0.014	0.223 $\pm$ 0.014
SBERT MPNet	110M	0.548 $\pm$ 0.015	0.338 $\pm$ 0.018	0.339 $\pm$ 0.011	<u>0.260</u> $\pm$ 0.013
Instructor Base	130M	0.303 $\pm$ 0.019	0.123 $\pm$ 0.017	0.158 $\pm$ 0.014	0.167 $\pm$ 0.011
SBERT Multilingual MPNet	280M	0.175 $\pm$ 0.022	0.098 $\pm$ 0.014	0.135 $\pm$ 0.016	0.143 $\pm$ 0.015
BGE	340M	<u>0.648</u> $\pm$ 0.019	<u>0.370</u> $\pm$ 0.015	<u>0.375</u> $\pm$ 0.014	0.256 $\pm$ 0.014
Instructor Large	370M	0.380 $\pm$ 0.017	0.198 $\pm$ 0.016	0.217 $\pm$ 0.011	0.227 $\pm$ 0.012
Qwen3 0.6B	600M	0.715 $\pm$ 0.010	0.459 $\pm$ 0.010	0.413 $\pm$ 0.009	0.326 $\pm$ 0.014
<i>Open-weights models ($\geq$1B)</i>					
Nemotron 1B	1.24B	0.839 $\pm$ 0.011	0.640 $\pm$ 0.012	0.541 $\pm$ 0.012	0.431 $\pm$ 0.018
Instructor XL	1.27B	0.524 $\pm$ 0.016	0.297 $\pm$ 0.019	0.291 $\pm$ 0.016	0.273 $\pm$ 0.014
Qwen3 4B	4.02B	0.832 $\pm$ 0.010	0.639 $\pm$ 0.012	0.567 $\pm$ 0.011	0.483 $\pm$ 0.016
INF-X-Retriever Pro	7.07B	0.846 $\pm$ 0.011	0.679 $\pm$ 0.009	0.598 $\pm$ 0.011	<u>0.516</u> $\pm$ 0.014
GritLM 7B	7.24B	<u>0.860</u> $\pm$ 0.008	<u>0.728</u> $\pm$ 0.012	0.605 $\pm$ 0.013	0.492 $\pm$ 0.016
Nemotron 8B	7.50B	0.897 $\pm$ 0.006	0.765 $\pm$ 0.006	0.654 $\pm$ 0.008	0.565 $\pm$ 0.014
Qwen3 8B	7.57B	0.859 $\pm$ 0.008	0.686 $\pm$ 0.010	0.603 $\pm$ 0.011	0.504 $\pm$ 0.014
KaLM Gemma3 12B	11.77B	0.827 $\pm$ 0.006	0.680 $\pm$ 0.009	<u>0.615</u> $\pm$ 0.007	0.505 $\pm$ 0.012

Table 2: Performance of different IR models on GUTENBENCH, considering 27 English translations and reported as nDCG@10. The best score is denoted in **bold** while the second best is underlined. In general, there is a consistent (negative) correlation between the query level and model performance, with L4 remaining the most challenging level for all competitive models.

6.2 Monolingual Performance (English)

We evaluate IR models using English queries to retrieve English documents, providing a controlled environment to analyze how models handle increasing retrieval complexity. To ensure robustness, we evaluate against a monolingual parallel corpus comprising 27 distinct English translations.

Table 2 shows retrieval performance declining as the taxonomy level increases: every competitive model scores highest on L1 and lowest on L4. Among sub-1B models, Qwen3 0.6B performs best across all levels. As expected, BM25 is competitive on L1, but degrades on L2–L4. Scaling beyond 1B parameters improves performance substantially, with Nemotron 8B achieving the best results across all four levels.

6.3 Cross-Lingual Performance

To understand the coverage of modern retrieval architectures beyond English, we apply our stratified taxonomy to a cross-lingual environment spanning 245 translations (140 unique languages). In this setting, each run pairs the English queries with one target-language translation of the corpus, so a model must retrieve documents written in a language other than that of the query. For this phase of the evaluation, strictly English-only models were naturally excluded. We categorize the target languages into high-resource, mid-resource, and low-resource tiers (refer to Appendix F for details) to

isolate how data availability impacts a retriever’s performance.

As illustrated in Figure 3, top multilingual models remain relatively robust in high-resource settings, with Nemotron 8B and Qwen3 8B showing only small degradation from English to mid-resource retrieval across levels. However, performance drops more sharply for low-resource languages, and this drop compounds with retrieval complexity. In low-resource settings, KaLM Gemma3 12B achieves the best score across all four levels.

Consequently, the intersection of high taxonomy level and low-resource languages results in a significant drop in performance for the tested IR models. This sharp decline highlights the current limitations of cross-lingual generalization, demonstrating that cross-lingual generalization remains a significant bottleneck for complex retrieval tasks in underrepresented languages.

6.4 Discussion

The experimental results provide empirical support for the proposed four-level taxonomy. Across nearly all evaluated models, retrieval performance declines as query complexity increases, with L4 consistently the weakest level for evaluated retriever. This trend holds in both monolingual and cross-lingual settings, suggesting that, indeed, the proposed taxonomy captures progressively more

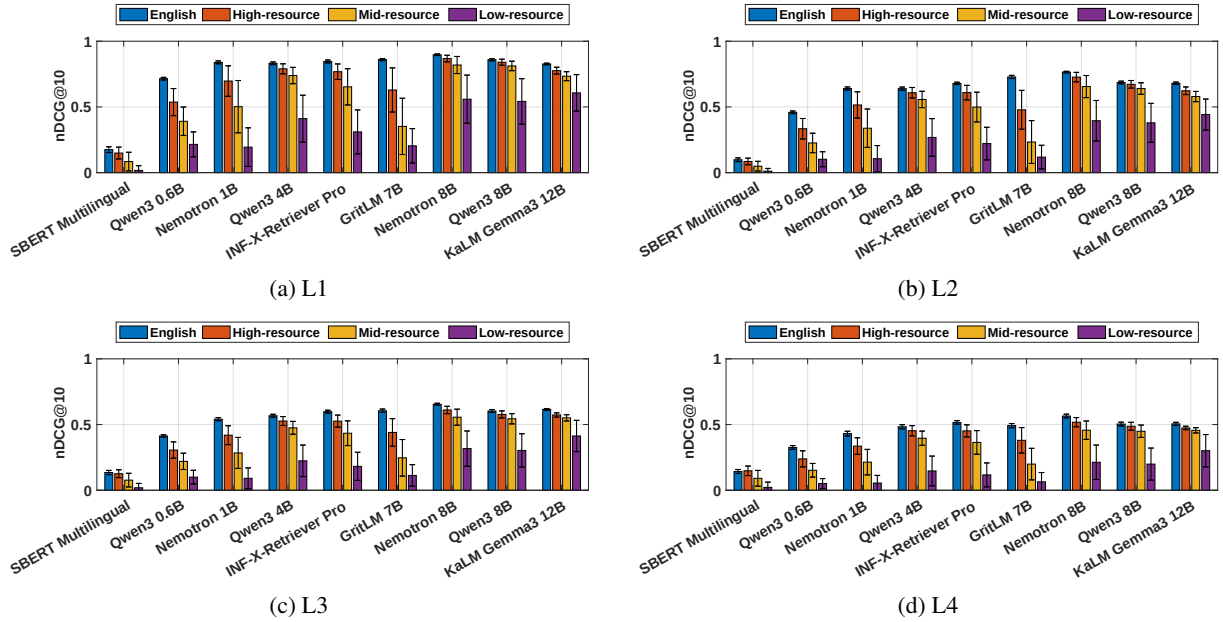


Figure 3: Retrieval performance across the four retrieval capabilities, comparing monolingual English retrieval with cross-lingual retrieval in high-resource, mid-resource, and low-resource languages. Performance degrades consistently as retrieval complexity increases, with the largest drops occurring for latent conceptual retrieval and low-resource cross-lingual settings, highlighting substantial limitations in current retrieval systems.

challenging retrieval behaviors.

The results also clarify the strengths and limitations of current retrieval architectures. BM25 remains competitive on L1, but degrades substantially on L2–L4. Its performance barely changes from L2 to L3, on the other hand, suggesting that some multi-document queries still contain explicit lexical evidence. Dense retrievers substantially improve absolute performance, especially at larger scales, but they preserve the same overall difficulty gradient rather than eliminating it. Still, L4 remains the clear bottleneck: even in English, the best model reaches only 0.565 nDCG@10, about 63% of its corresponding L1 score.

The cross-lingual experiments reveal that retrieval quality deteriorates systematically as language resource availability decreases. The combination of high retrieval complexity and low-resource languages produces the largest performance drops in our evaluation. There, KaLM Gemma3 12B achieves the best results across all four levels, showing that the strongest English retriever is not necessarily the most robust under low-resource transfer. GritLM 7B illustrates the converse, ranking near the top in English while degrading sharply across other languages.

Overall, these findings demonstrate the importance of fine-grained retrieval evaluation. Aggregate metrics can obscure capability differences

across retrieval behaviors, whereas GUTENBENCH enables a more precise diagnosis of retrieval strengths and failure modes across monolingual and cross-lingual settings.

7 Related Work

To position GUTENBENCH within the broader literature on retrieval evaluation, we review prior benchmarks and identify the gaps our framework addresses.

Zero-Shot and Multilingual Retrieval Benchmarks. Standardized benchmarks have significantly advanced the evaluation of information retrieval (IR) systems. Early unified frameworks like KILT (Petroni et al., 2021) established paradigms for evaluating retrieval-augmented pipelines. Subsequently, BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) became the standard for assessing zero-shot generalization across heterogeneous domains. More recent efforts have focused on massive cross-lingual scaling. Foundational datasets like mMARCO (Bonifacio et al., 2022) and CLIRMatrix (Sun and Duh, 2020) enabled deep cross-lingual evaluation. Most recently, the landscape has been defined by massive scaling and automation: AIR-Bench (Chen et al., 2025) introduced LLM-driven automated benchmark generation, NeuCLIR (Lawrie et al., 2025)

advanced rigorous cross-lingual test collections, and MMTEB (Enevoldsen et al., 2025) established a comprehensive ecosystem for massively multilingual embedding evaluation.

Reasoning-Intensive Retrieval. While many retrieval benchmarks do not explicitly separate lexical or semantic matching from cases where relevance depends on deeper inference, recent work has begun to target reasoning-intensive retrieval directly. Benchmarks such as BRIGHT (SU et al., 2025) and domain-specific suites like R2MED (Zhang et al., 2026) show that state-of-the-art dense retrievers often struggle when identifying relevant documents requires reasoning beyond surface-form or shallow semantic similarity.

Despite substantial progress in domain diversity, language coverage, and reasoning-intensive retrieval, existing benchmarks are generally organized around datasets, domains, languages, or task families, with results often summarized at those levels. This reporting granularity makes it difficult to isolate which retrieval capability failed and how that capability transfers across language-resource tiers. GUTENBENCH addresses this gap by combining a four-level retrieval taxonomy (lexical, semantic, explicit aggregation, latent conceptual) with evaluation across language-resource tiers, providing a highly controlled framework that isolates retrieval failure modes and prevents basic capabilities from masking deep conceptual deficits.

8 Conclusion

We introduced GUTENBENCH, a fine-grained dataset for information retrieval that disentangles retrieval competencies through a four-level taxonomy spanning lexical matching, semantic retrieval, explicit aggregation, and latent conceptual retrieval.

Our experiments show that retrieval performance degrades consistently as retrieval complexity increases, with the weakest performance on latent conceptual retrieval. While dense retrievers substantially outperform sparse lexical baselines, even the strongest models struggle when relevance depends on indirect conceptual associations rather than explicit semantic overlap. The results also show how naturally imbalanced query distributions can obscure these weaknesses, since many retrieval benchmarks are dominated by lexically transparent queries.

Future work will extend GUTENBENCH with unanswerable queries to study retrieval abstention: determining when retrievers should conclude that no supporting evidence exists in the corpus instead of returning misleading matches. This would enable evaluation not only of retrieval accuracy, but also of calibrated uncertainty and reliable “I do not know” behavior in RAG systems.

Limitations

While GUTENBENCH provides a rigorous, highly stratified evaluation framework, we acknowledge several limitations regarding its scope and methodology.

Document Length Constraints. To ensure precise, aligned cross-lingual evaluation, the corpus is segmented into short, non-overlapping chunks delimited by the NKJV section headings (Sec. 4.3). Therefore, GUTENBENCH strictly evaluates *passage* retrieval. It is not designed to test long-context document retrieval or extreme “needle-in-a-haystack” scenarios, which require different context-handling architectures.

Domain Specificity. GUTENBENCH operates exclusively on biblical text, so model performance on this corpus may not perfectly correlate with performance in other specialized domains, such as medical, financial, or legal retrieval. However, we use the Bible not as a representative domain, but as a diagnostic instrument: it provides a stable, compact, and massively parallel corpus that allows us to isolate retrieval mechanisms across many languages without prohibitive annotation costs. Thus, while absolute scores may be domain-dependent, the benchmark is designed to test retrieval behaviors that are relevant beyond this corpus.

Cross-Lingual Semantic Equivalence. Our multilingual evaluation relies on structural alignment, assuming that the semantic required to answer an English query is equally present in the corresponding translated chunk. However, translating ancient texts into other languages often involves distinct theological or linguistic choices, meaning the exact reasoning required for a L3 or L4 query might subtly shift across languages.

Corpora Distribution. GUTENBENCH distributes only the framework (queries, qrels, and evaluation scripts), not the underlying corpus text. To lower the barrier for adoption, we recommend

utilizing widely available public-domain translations (such as the King James Version or the World English Bible) for baseline evaluations. However, users remain responsible for legally obtaining any digital texts required to run the benchmark, particularly when evaluating modern, copyrighted cross-lingual editions.

Acknowledgements

The authors acknowledge support by the DFG through FOR 5359 (ID 459419731), TRR 375 (ID 511263698), SPP 2298 (IDs 441826958 and 441826958), and SPP 2331 (IDs 441958259, 553345933, and 466468799), by the Carl-Zeiss Foundation through the initiative AI-Care, and by the BMFTR award 01IS24071A.

The experiments were partly executed on the high performance cluster “Elwetritsch” at the RPTU Kaiserslautern-Landau which is part of the “Alliance of High Performance Computing Rheinland-Pfalz” (AHRP). We kindly acknowledge the support of RHRZ.

The authors thank the following individuals for their valuable support: Pastor Peterson Santos (PhD student, Andrews University, USA); Pastor Bruno Lopes (Saarland, Germany); and Pastor Ezequias Guimarães (São José dos Campos, Brazil).

References

- Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. [LLMs can be fooled into labelling a document as relevant: best café near me; this paper is perfectly relevant](#). In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024*, pages 32–41, Tokyo, Japan. Association for Computing Machinery.
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. [Llama-Embed-Nemotron-8B: A universal text embedding model for multilingual and cross-lingual tasks](#). *Preprint*, arXiv:2511.07025.
- Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. [mMARCO: A multilingual version of the MS MARCO passage ranking dataset](#). *Preprint*, arXiv:2108.13897.
- Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2025. [AIR-Bench: Automated heterogeneous information retrieval benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19991–20022, Vienna, Austria. Association for Computational Linguistics.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Sologatin, Ömer Veysel Çağatan, and 63 others. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. [Perspectives on large language models for relevance judgment](#). In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR ’23*, pages 39–50, Taipei, Taiwan. Association for Computing Machinery.
- Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. [On statistical bias in active learning: How and when to fix it](#). In *International Conference on Learning Representations*.
- Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. [Retrieving, rethinking and revising: The Chain-of-Verification can improve retrieval augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393, Miami, Florida, USA. Association for Computational Linguistics.
- Xinshuo Hu, Zifei Shan, Xinpeng Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. [KaLM-Embedding: Superior training data brings a stronger embedding model](#). *Preprint*, arXiv:2501.01028.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2).
- Kalervo Järvelin and Jaana Kekäläinen. 2002. [Cumulated gain-based evaluation of IR techniques](#). *ACM Trans. Inf. Syst.*, 20(4):422–446.
- Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. [Sufficient context: A new lens on retrieval augmented generation systems](#). In *The Thirteenth International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and](#)

- fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications.
- Dawn Lawrie, James Mayfield, Eugene Yang, Andrew Yates, Sean MacAvaney, Ronak Pradeep, Scott Miller, Paul McNamee, and Luca Soldani. 2025. [NeuCLIR-Bench: A modern evaluation collection for monolingual, cross-language, and multilingual information retrieval](#). *Preprint*, arXiv:2511.14758.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. [Latent retrieval for weakly supervised open domain question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
- 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, pages 9459–9474. Curran Associates, Inc.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. [Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations](#). In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, Virtual Event, Canada. Association for Computing Machinery.
- Robert (Munro) Monarch. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Manning Publications.
- Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2025. [Improving text embedding models with positive-aware hard-negative mining](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, pages 2169–2178, Seoul, Republic of Korea. Association for Computing Machinery.
- Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. [Generative representational instruction tuning](#). In *The Thirteenth International Conference on Learning Representations*.
- 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*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [KILT: a benchmark for knowledge intensive language tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Christopher Scialvolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. [Simple entity-centric questions challenge dense retrievers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. [One embedder, any task: Instruction-finetuned text embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121,

- Toronto, Canada. Association for Computational Linguistics.
- Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han Yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Serkan O Arik, Danqi Chen, and Tao Yu. 2025. **BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval**. In *The Thirteenth International Conference on Learning Representations*.
- Shuo Sun and Kevin Duh. 2020. **CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. **BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models**. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. **Large language models in medicine**. *Nature Medicine*, 29(8):1930–1940.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. **C-Pack: Packed resources for general Chinese embeddings**. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, pages 641–649, Washington DC, USA. Association for Computing Machinery.
- Yichen Yao, Jiahe Wan, Yuxin Hong, Mengna Zhang, Junhan Yang, Zhouyu Jiang, Qing Xu, Kuan Lu, Yinghui Xu, Wei Chu, Emma Wang, and Yuan Qi. 2025. **INF-X-Retriever: A pragmatic framework for reasoning-intensive dense retrieval**.
- Wan Zhang and Jing Zhang. 2025. **Hallucination mitigation for retrieval-augmented large language models: A review**. *Mathematics*, 13(5):856.
- Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. 2026. **R2MED: A benchmark for reasoning-driven medical retrieval**. *Preprint*, arXiv:2505.14558.
- Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. **MIRACL: A multilingual retrieval dataset covering 18 diverse languages**. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.
- 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*.
- Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2026. **KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model**. In *The Fourteenth International Conference on Learning Representations*.

A Getting Started

GUTENBENCH is released as two public artifacts: the dataset (queries, relevance judgments, and document boundaries), hosted on Hugging Face, and the evaluation software, hosted on GitHub. This appendix shows how to load the former and run the latter.

A.1 Loading the Dataset

We provide the dataset through Hugging Face at <https://huggingface.co/datasets/contemmmcm/gutenbench>. Releases are tagged, and the examples below pin the release evaluated in this paper (v2026). The dataset can be loaded using the Python programming language and the datasets library:

```
from datasets import load_dataset

topics = load_dataset(
    "contemmmcm/gutenbench",
    "topics",
    split="test",
    revision="v2026",
)
qrels = load_dataset(
    "contemmmcm/gutenbench",
    "qrels",
    split="test",
    revision="v2026",
)
documents = load_dataset(
    "contemmmcm/gutenbench",
    "documents",
    split="test",
    revision="v2026",
)
```

The documents configuration defines how the corpus is chunked into retrieval units: each record gives one document’s passage boundaries as OSIS references (Open Scripture Information Standard), which identify verses in book.chapter.verse form, so that GEN.1.1 denotes Genesis 1:1. For example:

```
{"doc_id": "doc0001",
```

```
"start": "GEN.1.1", "end": "GEN.2.7"}
{"doc_id": "doc0002",
 "start": "GEN.2.8", "end": "GEN.2.25"}
```

The corpus text itself is not distributed and must be paired with a legally obtained translation.

A.2 Running the Benchmark

The evaluation code is available at <https://github.com/contemmmcm/gutenbench> and can be installed as follows:

```
$ git clone \
  https://github.com/contemmmcm/gutenbench
$ cd gutenbench
$ conda create -n gutenbench python=3.11
$ conda activate gutenbench
$ conda install -c conda-forge openjdk=21 -y
$ pip install -r requirements.txt
```

After installation, place the retrieval corpus in the expected `bibles/` directory, as a verse-level CSV file in the format described in the repository. For example, the King James Version should be available as `bibles/KJV.csv`. The benchmark is then run through predefined make targets, where the first part specifies the retrieval model and the second part specifies the corpus:

```
$ make BM25/KJV
```

Each target downloads the queries and relevance judgments automatically from Hugging Face, chunks the corpus into the benchmark’s retrieval units, builds the index, runs the retrieval, and evaluates it with `trec_eval`, reporting MAP, nDCG@10, and Recall@100 for each taxonomy level. Alternatively, all evaluated retrievers can be executed on a given translation at once:

```
$ ./benchmark KJV
```

The use of make targets also keeps the evaluation pipeline modular, making it straightforward to add new retrievers without changing the core evaluation code.

B Queries Sources

The initial query pool was drawn from publicly available Bible-related web sources, including topical indexes, Bible-study materials, trivia collections, online QA pages, and video-based Bible QA. Table 3 summarizes representative sources and the query types they tend to contribute across the taxonomy. In addition, a small share of the queries (about 4% of the released set) was contributed directly by the maintainers from other sources, mainly to enrich the scarcer L3–L4 levels. These sources were used to collect candidate queries and, when

available, the supporting passages they propose, which enter the judgment pool only as additional candidates (Sec. 4.2). Relevance judgments were produced independently through the annotation pipeline described in Sec. 4.

C Assessor Qualification and Protocol Design

To design a reliable annotation protocol, we conducted an initial qualification study with a set of 50 golden queries carefully adjudicated, and used them to test potential assessors, both humans and LLMs. We binarized the graded judgments (grouping 1 and 2 as “Relevant”, and 0 as “Not Relevant”) to calculate Precision, Recall, and F1-scores. We set retention thresholds at a minimum F1-score of 0.85 and a minimum Recall of 0.90.

As a result, only human assessors with domain expertise and the LLMs met these thresholds. Among the retained assessors (humans and LLMs), inter-annotator agreement reached a Krippendorff’s α of **0.855**. This score is considered *reliable*, indicating that the agreement is systematic (Krippendorff, 2004).

Even so, Table 4 shows that LLMs consistently underperform the human experts. While top-performing models achieve near-human recall, their precision lags significantly. This precision gap is highly consequential for Retrieval-Augmented Generation (RAG). Typically, IR models are designed to favor recall, but RAG is somewhat sensitive to irrelevant context (false positives). Prone to a well-documented leniency in relevance labeling (Alaofi et al., 2024), LLM evaluators act as over-eager filters that admit tangentially related documents. Injecting these false positives into the context window introduces critical confounders, forcing the generator to synthesize irrelevant data and sharply increasing the risk of hallucinated responses.

These results directly shaped the annotation pipeline of Sec. 4. Because frontier LLMs approach human recall, we use them to propose candidate passages and to pre-fill relevance votes. Because their precision lags behind the experts, a qualified subject-matter expert validates or overrides every judgment, with particular attention to passages where the LLM votes disagree. The expert decision is authoritative. All underlying votes and judgments are preserved for auditing, and the released evaluation set contains only fully adjudi-

Source	URL	Reason to include
OpenBible Topics	https://www.openbible.info/topics/	Topical queries, mainly L3–L4.
GotQuestions	https://www.gotquestions.org/	Interpretive and doctrinal questions, mainly L2–L4.
StillFaith	https://stillfaith.com/	Bible-study and trivia-style Q&A, mainly L1–L2.
BibleStudy.org	https://www.biblestudy.org/	Passage-based and topical questions, mainly L1–L3.
Quora	https://www.quora.com/	Online Q&A, mainly L2–L4.
YouTube Bible Q&A channels	https://www.youtube.com/	Spoken Q&A, providing all kinds of queries (L1–L4).
King James Bible Online Trivia	https://www.kingjamesbibleonline.org/Bible-Trivia-Questions/	Trivia and factoid questions, mainly L1.
Logos Bible Trivia	https://www.logos.com/grow/100-bible-trivia-questions/	Curated factoid and passage-based questions, mainly L1–L2.
Biblical Hermeneutics Stack Exchange	https://hermeneutics.stackexchange.com/	Exegetical and passage-interpretation questions, mainly L2–L4.
Christianity Stack Exchange	https://christianity.stackexchange.com/	Doctrinal and comparative questions, mainly L3–L4.
BibleHub Questions	https://biblehub.com/q/index.html	Structured verse- and passage-focused questions, mainly L1–L2.
Blue Letter Bible FAQ	https://www.blueletterbible.org/faq/	FAQ-style topical and interpretive questions, mainly L2–L4.

Table 3: Examples of public sources used to construct or expand the candidate query pool. The taxonomy ranges indicate the kinds of queries each source tends to contribute more.

Assessor	Precision	Recall	F1
Human Expert	0.877	0.993	0.931
DeepSeek-V4 Pro	<u>0.832</u>	0.985	0.902
ChatGPT 5.5	0.821	0.978	0.893
Grok 4.2	0.842	0.941	0.889
Claude Opus 4.7	0.815	0.971	0.886
Gemini 3.1 Pro	0.806	0.949	0.872

Table 4: Assessor performance against the ground truth. While top LLMs achieve near-human recall, their lower precision reveals a persistent “positivity bias”.

cated judgments.

The taxonomy levels followed a similar protocol. Unlike relevance, assigning a level requires expertise in retrieval mechanisms rather than domain knowledge, particularly at the boundary between Levels 3 and 4. The levels were therefore assigned by the authors: given a query and its adjudicated relevant documents, LLM judges propose a taxonomy level, and the author may abide by or override their votes.

D Costs and Computational Overhead

This section outlines the resource requirements associated with GUTENBENCH, detailing both the estimated financial costs of the LLM-driven evidence pooling phase and the local computational overhead required for benchmark evaluation.

D.1 Initial Passages Pooling

Following the core idea of the TREC pooling methodology described in Sec. 4, we constructed an initial judgment pool for each query using five frontier zero-shot LLM-based systems. Table 5 details the average token consumption per model per 10 queries across each taxonomy level. As expected, average token consumption increases with retrieval complexity. This empirical scaling corroborates the taxonomy, suggesting that the higher-level queries (L3 and L4) impose a higher computational and reasoning overhead.

Level	Input Tokens	Output Tokens
L1	55,081	37,375
L2	59,361	43,303
L3	64,593	54,707
L4	84,975	95,053

Table 5: Average token consumption per model per 10 queries during the initial evidence pooling phase.

Because API pricing varies significantly depending on the specific model and provider used, we provide estimated costs based on standard OpenAI pricing at the time of publication as a baseline. The average financial cost for model **GPT-5.5** with high reasoning effort was:

- **L1:** \$1.34 per 10 queries
- **L2:** \$1.54 per 10 queries
- **L3:** \$1.90 per 10 queries

- **L4:** \$3.23 per 10 queries

D.2 Computing Resources

To provide a baseline for the computational overhead required to evaluate IR models locally on GUTENBENCH, we measured execution times using a single consumer-grade NVIDIA RTX 4090 (24GB VRAM). Table 6 reports the mean inference time required to evaluate Nemotron on a single translation on this hardware.

Model	Exec. Time	Seq. Size	Batch Size
Nemotron 1B	124 s	4096	2
Nemotron 8B	277 s	4096	1

Table 6: Mean execution time required to run GUTENBENCH on a single translation using a consumer-grade GPU (RTX 4090, 24GB).

These inference times highlight the practical advantage of GUTENBENCH’s compact corpus. By keeping the dataset small but highly stratified, the framework allows for rapid, local evaluation on consumer-grade hardware. While we do not propose GUTENBENCH as a replacement for exhaustive, heterogeneous benchmarks (such as BEIR or MTEB), its exceptionally low computational overhead makes it an ideal first-line diagnostic tool. For model developers and practitioners seeking to validate IR systems for specific target languages, it offers a fast, rigorous method to diagnose retrieval failure modes before committing to the higher costs of large-scale evaluations.

E Software

BM25 retrieval was implemented with Pyserini 1.4.0 (Lin et al., 2021). Dense retrievers were evaluated using Transformers 4.51.0 and Sentence-Transformers 5.3.0, with model-specific wrappers where required, including InstructorEmbedding 1.0.1, and GritLM 1.0.2. Query clustering for diversity sampling was implemented with UMAP (umap-learn 0.5.12) and HDBSCAN (hdbscan 0.8.44).

F Languages

We partition the non-English corpora into three resource categories: high, mid, and low, based on the linguistic diversity taxonomy established by Joshi et al. (2020). In our evaluation, high-resource languages correspond to their taxonomy Levels 4 and 5, mid-resource languages correspond to Level

3, and low-resource languages correspond to Levels 2, 1, and 0.⁶

F.1 High-Resource Languages (21)

Arabic (Standard), Catalan, Chinese (Simplified), Chinese (Traditional), Croatian, Dutch, Finnish, German, Hindi, Hindi (Roman script), Italian, Japanese, Korean, Polish, Portuguese (Brazil), Portuguese (Portugal), Serbian, Spanish, Spanish (Spain), Turkish, Vietnamese.

F.2 Mid-Resource Languages (15)

Afrikaans, Belarusian, Bengali (Bangla), Cebuano, Indonesian, Lithuanian, Malay, Modern Greek, Romanian, Tagalog, Tamil, Thai, Urdu, Urdu (Devanagari), Urdu (Roman).

F.3 Low-Resource Languages (104)

Acholi (Acoli), Albanian (Standard), Amharic, Armenian (Eastern), Assamese, Bari, Breton, Buli (Ghana), Cañar Highland, Quichua, Chimborazo Highland Quichua, Chinese (Cantonese), Chinese (Classical), Chuukese, Cree: Plains (nēhiyawēwin), Cusco Quechua, Datooga, Efate (North), Efik, Ekegusii, Esperanto, Fijian, Ganda, Gujarati, Haitian Creole, Hawaiian, Hiligaynon, Hiri Motu, Hmar, Hmong Daw, Hmong Njua, Huichol, Ibibio, Igbo, Ika, Iloko, Imbabura, Highland Quichua, Inuktitut: Eastern Arctic, Inuktitut: Labrador (Nunavimmiututut), Kalenjin, Kamano, Kamba (Kenya), Kannada, Kekchí, Khasi, Khmer, Kikuyu, Kinandi (Ndandi), Kosraean, Kurmanji Kurdish, Lamba, Lango (Uganda), Luba-Lulua, Lugbara, Lunda, Lushai, Luvale, Machame, Malayalam, Maori, Mara, Marathi, Mentawai, Mongo, Mongolian, Naga (Mzieme), Nepali, North Ndebele, Norwegian Bokmål, Nyanja, Nyankole, Nyoro, Obolo, Oriya, Panjabi, Pennsylvania German, Poqomchi’, Romani: Baltic, Romani: Kalderash, Romani: Ursari, Rundi, Samoan, Scots Gaelic, Setswana, Shona, Sindhi, Sinhala, Somali, Sotho, Northern, Southern Sotho, Swazi, Teso, Tetela, Tiv, Tonga (Nyasa), Tonga (Tonga Islands), Tonga (Zambia), Tsonga, Tzeltal, Umbundu, Urhobo, Welsh (Galés), Xhosa, Zande, Zulu.

G Volunteer Recruitment and Participation

All relevance assessors were recruited on a voluntary basis through online outreach to communities

⁶These language-resource classes are unrelated to our retrieval taxonomy levels L1–L4.

familiar with biblical and theological texts. Participation was fully unpaid, and volunteers were informed that their annotations would be used for research purposes in a publicly released benchmark.

The annotation interface was multilingual, allowing assessors to evaluate passages in their native languages. No personally identifying or sensitive information was collected beyond the annotations themselves, and assessors were free to stop or resume participation at any time.

H Instructions Given To Participants

The following instructions were shown to participants in the annotation interface before they assessed candidate passages for each query.

Guidelines

You are shown a Bible-related query and a list of candidate passages. For each passage, decide how much it helps answer that query.

The candidates were gathered automatically: some will be exactly right, some loosely related, and some will have nothing to do with the query. The task is to sort that out.

Your decisions will become the reference answer that search systems will be scored against. A system is penalised for either returning a passage you marked *Not relevant*, or for missing one you marked *relevant*.

Relevance scale:

- **Definitely relevant (2):** The passage directly and specifically addresses the query. It would be a top result for a user with this information need.
- **Partially relevant (1):** The passage is on topic and genuinely useful, but does not settle the query on its own. It backs up, illustrates, or supplies context for something a *Definitely relevant* passage states.
- **Not relevant (0):** The passage does not address the query, or its connection is only superficial.

The two mistakes to avoid:

1. **Missing real evidence.** Do not mark a passage *Not relevant* merely because the connection is not obvious at first glance. Ask instead: if this passage disappeared, would the answer lose real support? If it would, it is at least *Partially relevant*.
2. **Waving through the merely related.** Do not mark a passage relevant because it shares words, names, or a story with the query. Overlapping vocabulary is not support.

How to judge:

1. **Read the whole passage** before deciding.

2. **Judge the content, not your view of it.** Whether you agree with what the passage says, or think it important, has no bearing on the grade.
3. **Do not fill in what the passage does not say.** Judge what is there and what it directly supports, not what could be inferred with outside knowledge.
4. **Judge substance, not wording.** A passage can be *Definitely relevant* without sharing any vocabulary with the query, and *Not relevant* while sharing a great deal.

Cases that come up often:

1. **Only part of the query.** Grade the passage on the part it addresses. A passage that settles one half of a two-part query can be *Definitely relevant*.
2. **Needs its neighbours to make sense.** If the passage only supports the answer once you have read the surrounding verses, it is usually *Partially relevant* rather than *Definitely relevant*.
3. **Parallel accounts.** The same event or teaching often appears in several places. Each passage is judged on its own and they can all be relevant. Do not demote one merely because another says the same thing.
4. **You are unsure.** Choose the grade you would defend to a colleague, and say what made it hard.
5. **A relevant passage is missing.** The candidate list is not exhaustive. If you know a passage that belongs in the answer and it is not shown, add it.

I Bootstrap Stability Analysis

To assess whether the current evaluation set provides stable diagnostic estimates, we conducted a bootstrap stability analysis across all four taxonomy levels. The analysis is based on the English monolingual retrieval setting. With Nemotron 8B as the reference model, we computed the 95% confidence interval (CI) width of nDCG@10, with each query's score averaged over the English translations, as a function of the number of evaluated queries (N), drawing 10,000 bootstrap resamples at each step.

As illustrated in Figure 4, the CI width decreases rapidly as the number of sampled queries increases, followed by progressively smaller reductions. Although L1 and L2 contain much larger query pools, most of the variance reduction occurs in the early sampling regime, after which the curves begin to flatten.

The L4 subset contains only 103 queries, making it the smallest taxonomy level in the current release. We therefore do not claim that it exhausts

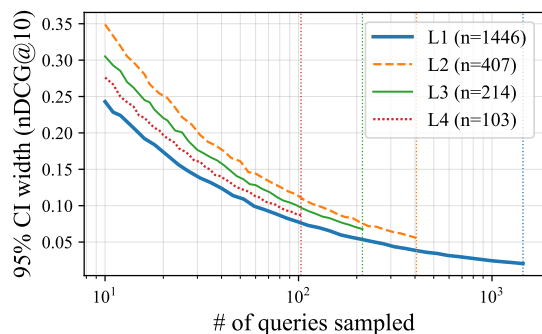


Figure 4: Empirical sample stability across all four retrieval levels on the English baseline. The x-axis is presented on a logarithmic scale to highlight the low-sample variance decay. The 95% Confidence Interval width demonstrates steep initial reductions before plateauing, showing diminishing returns for additional data collection.

the space of latent conceptual retrieval. However, its bootstrap curve follows the same broad convergence pattern as the other levels, suggesting that the current subset is sufficient for stable diagnostic estimates in the evaluated English setting. Additional L4 annotations would remain valuable for broader coverage, but the observed curve indicates diminishing returns for reducing uncertainty in the reported aggregate score.

This analysis supports the intended role of GUTENBENCH as a compact diagnostic benchmark: it is not designed to maximize dataset scale, but to provide reliable stratified signals about retrieval behavior under controlled evaluation conditions.

J Models and Instructions

Table 7 details the specific prompt templates and HuggingFace model identifiers used for the retrieval systems evaluated in this work. For models requiring designated query and document prefixes, we adhered to the formatting recommended by the developers. Models that require no special instructions, such as the SBERT family (Reimers and Gurevych, 2019), receive the raw query and document text and are omitted from the table.

Model	License	Hugging Face ID	Instructions
BGE (Xiao et al., 2024)	MIT	BAAI/bge-large-en-v1.5	Query: Represent this sentence for searching relevant passages: {query} Doc: {doc}
Instructor (Su et al., 2023)	Apache-2.0	hkunlp/instructor-base hkunlp/instructor-large hkunlp/instructor-xl	Query: Represent the question for retrieving relevant Bible passages: {query} Doc: Represent the Bible passage for retrieval: {doc}
INF-X-Retriever (Yao et al., 2025)	Apache-2.0	infly/inf-retriever-v1-1.5b infly/inf-retriever-v1 infly/inf-retriever-v1-pro	Query: Instruct: Given a Bible query, retrieve relevant passages that answer the query Query: {query} Doc: {doc}
GritLM 7B (Muennighoff et al., 2025)	Apache-2.0	GritLM/GritLM-7B	Query: < user > Given a Bible query, retrieve relevant passages that answer the query < embed > {query} Doc: {doc}
Qwen3 (Zhang et al., 2025)	Apache-2.0	Qwen/Qwen3-Embedding-0.6B Qwen/Qwen3-Embedding-4B Qwen/Qwen3-Embedding-8B	Query: Instruct: Given a Bible query, retrieve relevant passages that answer the query. {query} Doc: {doc}
Nemotron 1B (Moreira et al., 2025)	NVIDIA Open Model License. Llama 3.2 Community Model License Agreement.	nvidia/llama-nemotron-embed-1b-v2	Query: query: {query} Doc: passage: {doc}
Nemotron 8B (Babakhin et al., 2025)	NVIDIA License. Llama-3.1 Community License Agreement.	nvidia/llama-embed-nemotron-8b	Query: Instruct: Given a Bible question, retrieve passages that answer the question. {query} Doc: {doc}
KaLM Gemma3 12B (Hu et al., 2025; Zhao et al., 2026)	KaLM / Gemma terms + third-party terms.	tencent/KaLM-Embedding-Gemma3-12B-2511	Query: Instruct: Given a Bible query, retrieve documents that answer the query. Query: {query} Doc: {doc}

Table 7: Instructions and model identifiers used for each retrieval system.