C4: A Multilingual Benchmark for Retrieval-Augmented Generation based on the Catechism of the Catholic Church and its Compendium

Pius von Däniken, Mark Cieliebak, Jan Milan Deriu
Centre for Artificial Intelligence
ZHAW School of Engineering
{vode, ciel, deri}@zhaw.ch

Abstract

We introduce a new multilingual case study for evaluating retrieval augmented generation (RAG) systems, based on the *Catechism of the Catholic Church* (John Paul II, 1992b) and its *Compendium* (Benedict XVI, 2005a). The *Catechism* is a structured document with numbered paragraphs, officially translated into many languages under strict editorial alignment. The *Compendium* reformulates this material into a question-answer format with explicit citations to the corresponding paragraphs. Together, they form a set of parallel monolingual corpora that share identical semantic structure, enabling direct, controlled comparison of RAG performance across languages. Beyond its theological origin, this text pair closely mirrors real-world applications of RAG in institutional contexts, such as querying internal policy documents with associated FAQ-style summaries, making it a practical testbed for multilingual retrieval and grounded answer generation. We release our data collection scripts and baseline results for further research.

Keywords: retrieval augmented generation, benchmark, catechism

1. Introduction

Breakthrough advances in language technologies, driven by large language models (LLMs), have established retrieval-augmented generation (RAG) as a new paradigm for user interfaces in complex domains. In its basic form, a RAG system first retrieves relevant documents for a user query, then uses an LLM to generate an answer. In institutional contexts, RAG systems promise to make complex internal knowledge directly accessible through natural language interfaces. As such, the idea has gained wider popularity beyond academic research.

While the basic appeal of RAG systems lies in their apparent ease of construction by using off-the-shelf components for retrieval and generation, their evaluation is considerably harder, often requiring nuanced human judgment to assess relevance, faithfulness, and usefulness of the generated answers. This is particularly important in the case of domains that require deep expertise, such as legal or medical questions.

An ideal evaluation dataset should include high-quality, representative data for each stage of the process. User queries should mirror how people naturally seek information. Each query needs to be linked to all relevant passages required to form a complete response, allowing precise measurement of retrieval precision and recall. In addition, we need reliable gold-standard answers to evaluate generated responses. While information retrieval can be rigorously assessed given full relevance data, judging the quality of generated answers against reference texts remains challenging. Recent approaches have borrowed evaluation methods from summarization and LLM-as-a-judge frameworks.

We argue that the *Catechism of the Catholic Church* and its *Compendium* together meet these desiderata. The *Catechism* was compiled as "a sure and authentic reference text for teaching Catholic doctrine" and "a sure norm for teaching the faith" (John Paul II, 1992a). It contains 2,865 numbered paragraphs organized into four parts, subdivided into sections, chapters, and articles. The *Compendium* is described as "a faithful and sure synthesis of the Catechism of the Catholic Church" (Benedict XVI, 2005b). It reformulates the Catechism into 598 question-answer pairs, each linked to the corresponding paragraphs in the original text. Both texts are available in multiple languages. We show an example of a question-answer pair from the *Compendium* and its corresponding paragraphs from the *Catechism* in Table 1.

Structurally, these text pairs closely mirror the setup of a RAG system. The *Catechism* represents the knowledge base which should be made accessible. The questions from the *Compendium* can be considered queries to the knowledge base, and the citations provide a retrieval gold standard. Finally, a gold reference answer is also provided. Moreover, the data exists in multiple languages, allowing direct parallel comparisons for both monolingual and cross-lingual evaluation.

We therefore propose the *Catechism* and *Compendium* pair as a naturally occurring multilingual gold-standard dataset for RAG evaluation, which we refer to as *C4*. In this paper, we demonstrate

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7446–7456
11–16 May 2026.

its utility through multilingual baseline retrieval and generation experiments, and we make our scraping code and baseline results publicly available at https://github.com/vodezhaw/C4.

2. Related Work

Theological texts can serve as rich linguistic resources beyond their doctrinal content. The *Bible*, the most translated text in the world, has long been a valuable resource for computational linguistics (van Peursen, 2023). Numerous corpora derived from it exploit its multilingual coverage (e.g., Mayer and Cysouw, 2014; Bouma et al., 2020; Gueuwou et al., 2023) and have supported tasks such as cross-lingual information retrieval (Chew et al., 2006) and low-resource machine translation (Liu et al., 2021). In Arabic NLP, the *Hadith* literature has received considerable attention (Azmi et al., 2019) and recent work has even developed dedicated question answering corpora (Alnefaie et al., 2023).

Beyond academic research, public-facing applications have implemented conversational interfaces over the *Catechism*, such as the *CatéGPT* chatbot,[1] which demonstrates the practical appeal of retrieval-augmented question answering for theological texts.

The most common retrieval source for RAG question-answering datasets is Wikipedia (Wikipedia contributors, 2025). The *SQuAD* dataset (Rajpurkar et al., 2016), developed prior to the advent of transformer-based models (Vaswani et al., 2017), was designed to evaluate reading comprehension over individual Wikipedia passages. The *KILT* benchmark (Petroni et al., 2021) later consolidated multiple Wikipedia-based datasets under a unified snapshot, including *TriviaQA* (Joshi et al., 2017), *HotpotQA* (Yang et al., 2018), and *Natural Questions* (Kwiatkowski et al., 2019). Although these datasets were originally created before the notion of retrieval-augmented generation, they have become widely adopted for RAG evaluation.

Beyond English Wikipedia, *TyDi-QA* (Clark et al., 2020) introduced multilingual coverage by collecting question–answer pairs in 11 languages. Annotators were shown the first 100 characters of a Wikipedia article and asked to formulate a question, with the goal of eliciting realistic information-seeking behavior. They then used a search engine to find the most relevant article for the question and identified the passage containing the answer. Because *TyDi-QA* was built for a parallel monolingual setting—where retrieval is performed only within the same language as the query—*XOR-TyDi-QA* (Asai et al., 2021) extended it to a cross-lingual retrieval setting, in which questions posed in low-resource languages are answered using passages from English Wikipedia. Finally, *Mr. TyDi* (Zhang et al., 2021) further broadened the retrieval task by extending monolingual retrieval to the entirety of each language's Wikipedia, rather than a single article. *Mr. TyDi* relies on the original question–answer annotations from *TyDi-QA*. Building on this foundation, *MIRACL* (Zhang et al., 2023) extends the multilingual retrieval setting by adding additional relevance annotations for passages retrieved from other articles and expanding coverage to eight more languages.

Beyond Wikipedia-based benchmarks, other efforts have explored alternative retrieval domains. For instance, *xPQA* (Shen et al., 2023) addresses cross-lingual product question answering for online retail, and *WebFAQ* (Dinzinger et al., 2025) compiles large-scale FAQ data from diverse web sources.

Recent work has also introduced dedicated benchmarks and libraries for evaluating RAG systems more directly. *RAGAs* (Es et al., 2024) and *ARES* (Saad-Falcon et al., 2024) propose fully automated evaluation pipelines for context relevance, answer faithfulness, and answer relevance. *FaithEval* (Ming et al., 2025) and *NoMIRACL* (Thakur et al., 2024) focus on evaluation setups where retrieved evidence may be misleading or irrelevant. Complementary to such diagnostic evaluations, *RAG-Bench* (Friel et al., 2025) and *BERGEN* (Rau et al., 2024) provide unified frameworks for benchmarking retrieval and generation quality across multiple datasets. Chirkova et al. (2024) study multilingual RAG settings that involve full multilingual retrieval and generation.

Compared to these datasets and frameworks, *C4* provides a new naturally aligned, high-quality multilingual resource with explicit paragraph-level relevance annotations and corresponding gold-standard answers, complementing existing RAG evaluation resources.

3. Dataset

3.1. Data Source

Both the *Catechism* and the *Compendium* are publicly available in multiple languages on the Vatican's official online archive[2], in either PDF or HTML format. We focus on the languages that provide an HTML version following a common markup structure.

For the *Compendium*, we collected the English (en), French (fr), German (de), Hungarian (hu), Italian (it), Portuguese (pt), Slovenian (sl), and Spanish

[1] https://categpt.chat/, accessed 2025-10-23

[2] https://www.vatican.va/archive/ccc/index.htm, accessed 2025-10-22

	English	Español
Question	What is the importance of the Old Testament for Christians?	¿Qué importancia tiene el Antiguo Testamento para los cristianos?
Answer	Christians venerate the Old Testament as the true word of God. All of the books of the Old Testament are divinely inspired and retain a permanent value. They bear witness to the divine pedagogy of God's saving love. They are written, above all, to prepare for the coming of Christ the Savior of the universe.	Los cristianos veneran el Antiguo Testamento como verdadera Palabra de Dios: todos sus libros están divinamente inspirados y conservan un valor permanente, dan testimonio de la pedagogía divina del amor salvífico de Dios, y han sido escritos sobre todo para preparar la venida de Cristo Salvador del mundo.
CCC 121	The Old Testament is an indispensable part of Sacred Scripture. Its books are divinely inspired and retain a permanent value, for the Old Covenant has never been revoked.	El Antiguo Testamento es una parte de la sagrada Escritura de la que no se puede prescindir. Sus libros son divinamente inspirados y conservan un valor permanente (cf. DV 14), porque la Antigua Alianza no ha sido revocada.
CCC 122	Indeed, "the economy of the Old Testament was deliberately SO oriented that it should prepare for and declare in prophecy the coming of Christ, redeemer of all men." "Even though they contain matters imperfect and provisional, The books of the Old Testament bear witness to the whole divine pedagogy of God's saving love: these writings "are a storehouse of sublime teaching on God and of sound wisdom on human life, as well as a wonderful treasury of prayers; in them, too, the mystery of our salvation is present in a hidden way."	En efecto, «el fin principal de la economía del Antiguo Testamento era preparar la venida de Cristo, redentor universal». «Aunque contienen elementos imperfectos y pasajeros», los libros del Antiguo Testamento dan testimonio de toda la divina pedagogía del amor salvífico de Dios: «Contienen enseñanzas sublimes sobre Dios y una sabiduría salvadora acerca de la vida del hombre, encierran admirables tesoros de oración, y en ellos se esconden el misterio de nuestra salvación» (DV 15).
CCC 123	Christians venerate the Old Testament as true Word of God. The Church has always vigorously opposed the idea of rejecting the Old Testament under the pretext that the New has rendered it void (Marcionism).	Los cristianos veneran el Antiguo Testamento como verdadera Palabra de Dios. La Iglesia ha rechazado siempre vigorosamente la idea de prescindir del Antiguo Testamento so pretexto de que el Nuevo lo habría hecho caduco (marcionismo).

Table 1: Example question–answer pair number 21 from the *Compendium* (Benedict XVI, 2005a), and paragraphs 121–123 cited from the *Catechism* (John Paul II, 1992b), shown in English and Spanish translation.

(es) versions. Each language version is published as a single HTML file, composed primarily of <p> elements. Questions appear within a <b> tag inside a <p> element and are prefixed by their question number. Each question is followed by a list of paragraph references to the Catechism, and then by one or more paragraphs containing the answer.[3] Minor formatting inconsistencies were addressed through exception rules in the scraping script. After extracting the Compendium in all languages, we harmonized the paragraph reference lists across versions: in cases where a language omitted or added references, we adopted the reference set used by the majority of languages.

For the Catechism, we collected the English (en), French (fr), German (de), Italian (it), Spanish (es), and Latin (la) versions. Each version provides an index page linking to multiple section files that contain the paragraphs. These links follow a consistent pattern, allowing sequential scraping. Each paragraph begins with its number (within a <p> tag),

[3]We excluded Romanian and Swedish, as their HTML versions include the paragraph references after the answer.

which enables ordered retrieval.[4]

Since the texts are under copyright by the Libreria Editrice Vaticana, we are not permitted to redistribute the scraped data. However, to ensure reproducibility, we make our scraping and preprocessing scripts publicly available[5], allowing others to recreate the dataset directly from the official Vatican sources.

3.2. Data Structure and Alignment

The final dataset consists of two components: the *Compendium* question-answer pairs and the *Catechism* paragraphs.

Each *Compendium* is represented as a list of *(question_num, question_text, answer_text, paragraph_refs)* where *question_num* is the numeric identifier, *question_text* and *answer_text* contain the question and its corresponding answer, and *paragraph_refs* is a list of paragraph numbers referring to paragraphs in the *Catechism*. Since the list of paragraph references is harmonized across all languages, the same question number refers to the same paragraphs in all languages.

The *Catechism* itself is stored as a list of *(paragraph_num, paragraph_text)* pairs, where each paragraph is identified by its canonical number.

Because all *Compendium* and *Catechism* language versions share identical numbering and reference structures, the dataset is perfectly aligned across languages. This alignment enables direct multilingual comparison and supports cross-lingual retrieval and evaluation. The *Catechism* contains 2,865 paragraphs, and the *Compendium* comprises 598 question–answer pairs. Each question–answer pair references between 1 and 25 paragraphs, with an average of 4.8 and a median of 4. In total, 562 out of the 598 pairs reference 10 paragraphs or fewer.

4. Experiments

To demonstrate the utility of *C4*, we conduct two representative experiments: one evaluating retrieval and one evaluating answer generation, each in both monolingual and cross-lingual settings.

4.1. Retrieval Evaluation

We first evaluate the dataset in a cross-lingual retrieval setting. Each *Compendium* language (queries) is paired with each *Catechism* language (retrieval base), allowing both monolingual and cross-lingual retrieval performance to be assessed across all available languages.

We benchmark four retrieval models, one sparse and three dense, using *Pyserini*[6] (Lin et al., 2021). As the sparse baseline, we employ *BM25* (Robertson and Walker, 1994) with default parameters. As dense retrievers, we use *BGE-M3*[7] (Chen et al., 2024), *mpnet*[8] (Reimers and Gurevych, 2020), and *MiniLM*[9] (Reimers and Gurevych, 2019). We select *BGE-M3* and *mpnet* for their strong performance on the multilingual retrieval portion of the *MTEB* benchmark[10] (Muennighoff et al., 2023) and their compatibility with *Pyserini*. *MiniLM* is included due to its widespread use and speed, allowing us to assess how a monolingual embedding performs in this setting.

For each *Compendium* question, the retriever ranks all *Catechism* paragraphs according to their similarity to the question text. The gold paragraphs are those referenced by the *Compendium* for the corresponding question. Retrieval effectiveness is evaluated using Recall@10, which measures the proportion of all relevant documents that appear within the top 10 retrieved results. The cutoff of 10 matches the number of context paragraphs provided to the generator in the subsequent answer generation experiments (see Section 4.2).

Table 2 reports average Recall@10 scores for each retriever and language pair. As expected, *BM25* shows very limited cross-lingual effectiveness, with only typologically similar pairs such as pt–es achieving modest recall (0.261). It also underperforms dense retrievers even in monolingual settings, surpassing only the English-only *MiniLM* on non-English queries. *MiniLM* performs competitively only for the en–en pair (Recall@10 = 0.453). *mpnet* likewise peaks on en–en retrieval but yields substantially better results than *MiniLM* across other monolingual and cross-lingual configurations. The best overall results are obtained with *BGE-M3*, whose monolingual recall ranges from 0.509 (de–de) to 0.531 (it–it), and whose lowest cross-lingual score is 0.393 (sl–la). This model demonstrates the strongest cross-lingual retrieval capabilities among all tested systems. Despite these improvements, there remains significant

[4] Although a Portuguese HTML version exists, it contains extensive formatting errors and was therefore excluded.

[5] https://github.com/vodezhaw/C4/

[6] https://github.com/castorini/pyserini, accessed 2025-10-22

[7] https://huggingface.co/BAAI/bge-m3, accessed 2025-10-22

[8] sentence-transformers/paraphrase-multilingual-mpnet-base-v2, accessed 2025-10-22

[9] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, accessed 2025-10-22

[10] https://huggingface.co/spaces/mteb/leaderboard, accessed 2025-10-22

Retriever	Query	de	en	es	fr	it	la
BM25	de	0.359	0.093	0.068	0.067	0.106	0.046
	en	0.054	0.330	0.115	0.137	0.100	0.063
	es	0.069	0.137	0.364	0.113	0.186	0.051
	fr	0.027	0.148	0.095	0.365	0.101	0.050
	it	0.056	0.124	0.130	0.128	0.380	0.076
	hu	0.031	0.071	0.075	0.038	0.083	0.021
	pt	0.058	0.124	0.261	0.100	0.171	0.058
	sl	0.021	0.024	0.013	0.015	0.027	0.027
BGE-M3	de	0.509	0.472	0.466	0.459	0.462	0.413
	en	0.450	0.523	0.504	0.519	0.506	0.461
	es	0.455	0.512	0.530	0.506	0.510	0.482
	fr	0.441	0.501	0.511	0.522	0.509	0.460
	it	0.460	0.508	0.518	0.512	0.531	0.480
	hu	0.423	0.450	0.449	0.445	0.449	0.399
	pt	0.444	0.507	0.519	0.512	0.505	0.475
	sl	0.418	0.435	0.437	0.432	0.433	0.393
mpnet	de	0.378	0.386	0.378	0.367	0.380	0.110
	en	0.384	0.461	0.435	0.422	0.440	0.141
	es	0.364	0.425	0.432	0.406	0.426	0.132
	fr	0.363	0.403	0.401	0.401	0.410	0.129
	it	0.359	0.419	0.415	0.399	0.425	0.127
	hu	0.336	0.360	0.358	0.349	0.366	0.101
	pt	0.362	0.424	0.426	0.401	0.426	0.129
	sl	0.330	0.354	0.337	0.332	0.344	0.104
MiniLM	de	0.224	0.052	0.025	0.042	0.020	0.034
	en	0.053	0.453	0.083	0.157	0.083	0.119
	es	0.027	0.085	0.217	0.069	0.097	0.066
	fr	0.045	0.142	0.079	0.288	0.079	0.084
	it	0.020	0.067	0.087	0.058	0.252	0.085
	hu	0.011	0.015	0.013	0.013	0.011	0.012
	pt	0.024	0.079	0.159	0.077	0.095	0.075
	sl	0.013	0.013	0.010	0.011	0.011	0.007

Table 2: Recall@10 results of the multilingual retrieval experiments. *Query* indicates the language of the *Compendium* used, and the columns indicate the language of the *Catechism*.

headroom: even in the best monolingual cases, only about half of the relevant paragraphs are retrieved. However, since 94% of queries contain fewer than ten relevant documents, Recall@10 = 1 is theoretically attainable for the vast majority of cases.

4.2. Answer Generation

We now evaluate the dataset's utility for answer generation. We consider all monolingual settings, as well as selected cross-lingual settings for languages that have a *Compendium* but no *Catechism*. Specifically, we include hu–en, pt–es, sl–en, and en–la.

We test three retrieval settings: *No Retrieval*, where an LLM is directly prompted to answer without any context; *BGE-M3*, where the model receives the top 10 paragraphs retrieved by *BGE-M3*; and *Oracle*, where all gold paragraphs are provided as context. In all cases, the question is presented in the *Compendium* language, and the model is expected to respond in the same language, while the context is passed in the *Catechism* language.

We evaluate two instruction-tuned LLMs: *Llama-3.2-3B-Instruct* (*LL3.2-3B*)[11] and *gemma-3-4b-it* (*G3-4B*)[12] (Gemma Team et al., 2025). We use two different system prompts. The first is for the *No Retrieval* setting:

System prompt, no retrieval

[11] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, accessed 2025-10-22

[12] https://huggingface.co/google/gemma-3-4b-it, accessed 2025-10-22

Q	R	Gen	No Retrieval			bge-m3			Oracle		
			R-1	R-2	CEv.	R-1	R-2	CEv.	R-1	R-2	CEv.
en	en	G3-4B	0.229	0.049	0.515	0.292	0.092	0.532	0.316	0.103	0.666
		LL3.2-3B	0.241	0.064	0.442	0.324	0.129	0.554	0.345	0.142	0.645
fr	fr	G3-4B	0.228	0.050	0.403	0.288	0.088	0.510	0.314	0.103	0.644
		LL3.2-3B	0.232	0.052	0.299	0.304	0.109	0.525	0.325	0.125	0.596
de	de	G3-4B	0.210	0.032	0.366	0.268	0.068	0.454	0.298	0.085	0.580
		LL3.2-3B	0.211	0.038	0.250	0.303	0.108	0.491	0.328	0.123	0.558
it	it	G3-4B	0.214	0.039	0.429	0.268	0.077	0.521	0.296	0.093	0.638
		LL3.2-3B	0.216	0.043	0.302	0.301	0.115	0.532	0.332	0.134	0.596
es	es	G3-4B	0.254	0.068	0.418	0.308	0.109	0.515	0.337	0.127	0.616
		LL3.2-3B	0.261	0.075	0.335	0.347	0.148	0.530	0.368	0.163	0.593
hu	en	G3-4B	0.139	0.027	0.299	0.163	0.034	0.341	0.179	0.041	0.451
		LL3.2-3B	0.150	0.029	0.181	0.172	0.037	0.304	0.181	0.038	0.367
pt	es	G3-4B	0.220	0.040	0.424	0.268	0.069	0.522	0.294	0.082	0.644
		LL3.2-3B	0.224	0.042	0.301	0.279	0.086	0.509	0.302	0.100	0.574
sl	en	G3-4B	0.142	0.017	0.293	0.164	0.028	0.370	0.187	0.038	0.504
		LL3.2-3B	0.141	0.018	0.121	0.157	0.026	0.242	0.173	0.032	0.299
en	la	G3-4B	0.228	0.049	0.519	0.270	0.075	0.461	0.303	0.094	0.610
		LL3.2-3B	0.240	0.065	0.439	0.293	0.094	0.474	0.307	0.104	0.563

Table 3: Rouge-1 (R-1), Rouge-2 (R-2), and CheckEval (CEv.) scores for all combinations of query language (Q), retrieval language (R), LLM generator (Gen), and context setting.

> You are a helpful assistant knowledgeable about catholic theology and the Catechism of the Catholic Church.
> The user question is in {{ question_language }}. Respond **only** in {{ question_language }}.
> Answer the question based on catholic doctrine and your knowledge of the Catechism.
> Your answer should target the laity and be concise, spanning at most 6 sentences.
> If you do not know an answer, explicitly state that you don't know.

We use a second system prompt for the RAG setting:

> **System prompt, RAG**
> You are a helpful assistant knowledgeable about catholic theology and the Catechism of the Catholic Church.
> Your task is to answer questions based on paragraphs from the Catechism that are provided by the user.
>
> The paragraphs will be provided in {{ context_language }}.
> The user question is in {{ question_language }}. Respond **only** in {{ question_language }}.
>
> Your answer should target the laity and be concise, spanning at most 6 sentences.
> Base your answer strictly on the provided paragraphs. If they do not contain the relevant information, explicitly state that you don't know.
> Do not include your own opinions or interpretations beyond what is supported by the text.

In the user prompt we provide the retrieved passages and user query.

Generated answers are compared against the gold reference answers from our dataset. We report Rouge-1 and Rouge-2 (Lin, 2004), as well as an LLM-as-a-judge evaluation inspired by *CheckEval* (Lee et al., 2025). The key idea of *CheckEval* is to decompose qualitative evaluation into a set of binary checklist questions, such that a high-quality answer yields many "yes" responses. We adopt ten such questions, grouped by correctness, completeness, and focus (see Table 4). Each answer receives a normalized score (i.e., the number of positive responses divided by 10), and we report the average across all examples. To ensure a strong and independent evaluator and to minimize self-evaluation bias (Wataoka et al., 2024), we use `gpt-5-nano`[13] as the judging model.

Table 3 presents the results of the answer generation evaluation. As expected, all metrics improve with increasing quality of retrieved context.

[13]Model version: gpt-5-nano-2025-08-07

Correctness	
1	Does the generated answer express the same essential doctrinal meaning as the canonical answer?
2	Does the generated answer avoid any statement that contradicts the canonical answer?
3	Are key terms and concepts used with the same meaning and nuance?
Completeness	
4	Does the generated answer include all major elements present in the canonical answer?
5	Does the generated answer provide sufficient explanation for the key ideas?
6	Is the emphasis of the generated answer roughly proportional to the canonical answer?
Focus	
7	Does the generated answer directly answer the question asked?
8	Does the generated answer avoid tangential or unrelated information?
9	Does the generated answer stay within the canonical answer's scope (no speculative additions)?
10	Is the generated answer appropriately concise?

Table 4: Binary questions used for CheckEval-style evaluation of generated answers compared to gold answers.

The highest overall ROUGE-1 (0.368) and ROUGE-2 (0.163) scores are achieved by *LL3.2-3B* in the monolingual Spanish setting with *Oracle* retrieval, while the best *CheckEval* score (0.666) is obtained by *G3-4B* in monolingual English under the same retrieval condition. Although *LL3.2-3B* generally attains higher ROUGE scores than *G3-4B* across most settings, the trend reverses for *CheckEval*, where *G3-4B* performs slightly better. The strongest cross-lingual results are observed for Portuguese queries with Spanish *Oracle* retrieval, yielding a ROUGE-1 of 0.302 for *LL3.2-3B* and a *CheckEval* score of 0.644 for *G3-4B*. However, given known limitations of LLM-based evaluation methods (e.g., Wataoka et al., 2024; Shi et al., 2025), these score differences should be interpreted with caution, and minor gaps between models or languages may not be statistically meaningful.

5. Discussion

The experiments presented here are intended as demonstrations of how the dataset can be used and to establish baseline results. The parallel structure of the *Catechism–Compendium* pair enables controlled evaluation of retrieval and generation across multiple languages using the same underlying content. The reported results therefore serve as a starting point for future work.

The retrieval experiments show that even strong multilingual dense retrievers exhibit variation across languages. Particularly relevant are cross-lingual settings where no retrieval base exists for the query language. While a common practical solution is to retrieve from a high-resource language such as English, our results suggest that, under comparable conditions, retrieving from a typologically similar language can be advantageous. Because the dataset provides clear ground-truth mappings across languages, future work can examine these effects more systematically.

The answer generation experiments illustrate how retrieval quality affects grounded generation across languages. Since the source texts are publicly available, it is likely that they are included in the pretraining data of many large language models. The *No Retrieval* setting thus reflects baseline generation performance, while the *Oracle* setting defines an upper bound, as the dataset provides exhaustive relevance annotations rather than sparse judgments.

Overall, these experiments establish initial baselines for evaluating retrieval-augmented generation in multilingual settings, allowing retrieval and generation to be assessed both independently and end-to-end.

6. Conclusion

We introduced *C4*, a multilingual benchmark for retrieval-augmented generation (RAG) derived from the *Catechism of the Catholic Church* and its *Compendium*. Owing to the texts' close alignment across languages, it enables controlled evaluation of retrieval and generation in both monolingual and cross-lingual settings. Our baseline experiments demonstrate the dataset's utility for examining the strengths and limitations of current multilingual retrievers and generators. We release our scraping code and invite the community to include *C4* among the benchmarks used to evaluate RAG systems.

7. Limitations

C4 is a relatively small, domain-specific dataset, consisting of 2,865 paragraphs for retrieval and 598 question–answer pairs. Its limited size is offset by the high quality of the material, its natural alignment, and its availability in multiple languages.

C4 currently covers only European languages, predominantly Romance ones. Future work could expand coverage by parsing additional PDF versions available in Arabic, Chinese, and other languages.

Because the questions in the *Compendium* were created through editorial reformulation rather than

collected from natural user queries, they do not fully meet the desideratum of natural query formulation (see Section 1). The dataset's domain specificity may further limit the generalizability of findings.

Our baseline experiments employ comparatively simple models: standard multilingual dense retrievers and compact instruction-tuned generators (3B and 4B parameters). These choices provide accessible baselines rather than state-of-the-art results. Moreover, as the texts are publicly available and likely included in large language model pretraining data, generation results may partially reflect memorized knowledge rather than retrieval effects.

Finally, we do not provide human evaluation of the generation experiments but rely on fully automated metrics. We acknowledge the limitations of LLM-based evaluation methods.

8. Ethical Considerations

This work uses publicly available texts published by the Vatican. Both the *Catechism of the Catholic Church* and its *Compendium* are protected by the copyright of the Libreria Editrice Vaticana. To respect these rights, we do not redistribute any textual material; instead, we release only our scraping and preprocessing scripts that allow others to reproduce the dataset from the official sources.

The texts contain doctrinal statements reflecting the specific theological and cultural context of the Roman Catholic Church, including positions on sensitive topics. The dataset is intended and used solely as a structured multilingual resource for evaluating retrieval-augmented generation. We do not take any position on the theological or moral content of the texts and emphasize that the data should be used respectfully and within appropriate research contexts.

Since the texts are sourced from the web, they are likely included in the pretraining data of large language models used for the generation step, and this should be considered when interpreting evaluation results.

9. Acknowledgements

This work was supported by the Swiss National Science Foundation (SNF) within the project "Unified Model for Evaluation of Text Generation Systems (UniVal)" [200020_219819].

10. Bibliographical References

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 90–97, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Aqil M. Azmi, Abdulaziz O. Al-Qabbany, and Amir Hussain. 2019. Computational and natural language processing based studies of hadith literature: a survey. *Artificial Intelligence Review*, 52(2):1369–1414.

Benedict XVI. 2005a. *Compendium of the Catechism of the Catholic Church*, english translation. The Holy See: Vatican.va, English version. Promulgated by Benedict XVI; accessed 2025-10-22.

Benedict XVI. 2005b. Motu proprio *For the Approval and Publication of the Compendium of the Catechism of the Catholic Church*. The Holy See: Vatican.va, English version. Promulgated by Benedict XVI; accessed 2025-10-22.

Gerlof Bouma, Evie Coussé, Trude Dijkstra, and Nicoline van der Sijs. 2020. The EDGeS diachronic Bible corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5232–5239, Marseille, France. European Language Resources Association.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. 2006. Evaluation of the Bible as a resource for cross-language information retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74, Sydney, Australia. Association for Computational Linguistics.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vas-

silina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. Webfaq: A multilingual collection of natural qa datasets for dense retrieval.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

John Paul II. 1992a. Apostolic constitution *Fidei Depositum*: On the publication of the catechism of the catholic church. The Holy See: Vat-

ican.va, English version. Promulgated by John Paul II; accessed 2025-10-22.

John Paul II. 1992b. *Catechism of the Catholic Church*, english translation. The Holy See: Vatican.va, English version. Promulgated by John Paul II; accessed 2025-10-22.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of Bibles in low-resource machine translation. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *International Conference on Representation Learning*, volume 2025, pages 29430–29456.

Niklas Muennighoff, Nouamane Tazi, Loic 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.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual

using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. ACM/Springer.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. 2023. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge.

Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

Willem van Peursen. 2023. *Computational Linguistic Analysis of the Biblical Text*, Semitic Languages and Cultures, pages 223–272. Open Book Publishers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Wikipedia contributors. 2025. Wikipedia, The Free Encyclopedia. Accessed: October 22, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.