bLLeQA: Benchmarking LLMs for Grounded Legal Question-Answering in French and Dutch

Nikolay Banar*, **Ehsan Lotfi***, **Jens Van Nooten**, **Marija Kliocaite**, **Walter Daelemans**

University of Antwerp, Belgium

Correspondence: nicolae.banari@uantwerpen.be

Abstract

Retrieval-augmented generation (RAG) systems can play an important role in making law more accessible. However, large and reliable resources for training and benchmarking such systems remain scarce, especially for under-resourced languages like Dutch. To address this gap, and building on previous work (Louis et al., 2024), we introduce bLLeQA, a bilingual parallel question-answering dataset grounded in Belgian legal resources, both in French and Dutch. The dataset contains aligned questions, answers, and supporting articles in both languages, enabling evaluation of both retrieval and end-to-end RAG pipelines. Using bLLeQA, we benchmark the full RAG pipeline in a zero-shot setting, covering retrieval, citation extraction, refusal behavior, and generation quality. Our experiments show that open-weight models are competitive with proprietary models in retrieval and citation extraction, but lag behind in generation quality in the RAG pipeline. Across all models, refusal capability remains weak, meaning that models do not reliably detect when the provided supporting sources are incomplete. In addition, the end-to-end RAG setup still yields a substantial share of flawed responses, reaching 20% even in the best-case scenario.

1 Introduction

Access to justice remains a critical challenge for individuals and communities around the world. Disagreements with landlords, workplace disputes, and other civil justice problems are common occurrences, yet a significant portion of the population lacks the knowledge to resolve these issues or understand their rights (Balmer et al., 2010). The primary barriers to accessing justice are the prohibitive costs of legal counsel and a widespread lack of awareness about available legal options (Redelaar et al., 2024). Automated legal question answering (LQA) systems represent a promising avenue to democratize access to legal information by providing affordable, scalable assistance to broad audiences (Redelaar et al., 2024). LQA involves responding to queries, a task traditionally performed by domain experts, by reviewing relevant laws, interpreting statutes, and applying legal principles to specific facts (Ariai et al., 2025).

The rapid evolution of large language models (LLMs) has significantly advanced the capabilities of LQA systems. These models demonstrate remarkable proficiency in processing large volumes of text and generating human-like responses, offering opportunities to streamline legal research for professionals and lower barriers to information for the general public (Akarajaradwong et al., 2025; He et al., 2026). However, the application of LLMs in the legal domain is not without challenges. Legal queries require high precision, and general-purpose LLMs are prone to hallucinations and relying on outdated information. To mitigate this, retrieval-augmented generation (RAG) has become a standard framework for reliable LQA. By retrieving authoritative legal sources to ground the model's generation, RAG systems ensure that answers are verifiable, auditable, and based on up-to-date legal texts (He et al., 2026).

Despite the proliferation of LQA systems, there is a significant disparity in their availability across different languages and jurisdictions. Most existing resources and benchmarks focus on resource-rich languages such as English and common law traditions. However, legal systems are inherently jurisdiction-specific; a digital legal aid system designed for the US is ineffective for a civil law country operating under a different framework and language, and there is a critical need for LQA solutions that can handle the nuances of local law (Redelaar et al., 2024).

As a multilingual country, Belgium invests significant resources to consolidate its laws in both

*indicates equal contribution

French and Dutch, using qualified legal professionals. This results in a highly valuable resource for research in multilingual legal applications. Leveraging this potential and building on the Belgian LLeQA dataset (Louis et al., 2024) in French, we introduce the bilingual LLeQA (bLLeQA) that extends LLeQA to a parallel French-Dutch setting, via alignment, translation, and refinement of the original annotations in collaboration with a legal professional. Using bLLeQA, we conducted extensive benchmarking of LLMs in three RAG scenarios, providing insights into the performance of different LLMs in handling legal queries and citations. Our contributions are the following:

- We create and publish a parallel bilingual dataset for retrieval-based legal question-answering in French and Dutch, based on Belgian legislation.[1]
- We benchmark a wide range of open and proprietary LLMs on legal question-answering in French and Dutch, under different RAG settings.[2]

2 Related Work

Existing resources for legal question answering can be categorized by the scope and complexity of the task, ranging from document retrieval to answer generation and full RAG evaluation.

Legal Retrieval Legal document retrieval concerns the task of finding and ranking documents relevant to a given query from a large set of candidates. It is commonly studied over a corpus of cases (legal case retrieval) or articles (statutory article retrieval). In case retrieval, LeCaRD (Ma et al., 2021) (later extended as LeCaRDv2 (Li et al., 2024)) provided one of the first resources for the Chinese legal system, while ECtHR-PCR (T.y.s.s. et al., 2024) and CLERC (Hou et al., 2025) offered similar datasets for European and US jurisdictions, respectively. In statutory article retrieval, examples include BSARD (Louis and Spanakis, 2022), bBSARD (Lotfi et al., 2025b), and STARD (Su et al., 2024), which provided sizable resources based on Belgian and Chinese legislation in French, Dutch, and Chinese. In these datasets, each legal query is labeled with a set of articles that human annotators identified as necessary to answer the query.

[1] https://huggingface.co/datasets/clips/bLLeQA
[2] https://github.com/nikolay-banar/blleqa

LQA LQA is the task of providing valid responses to legal queries, with or without having access to legal documents (open- or closed-book). The response can be extracted from a given context (CJRC (Duan et al., 2019), EQUALS (Chen et al., 2023)), selected from provided options or choices (JEC-QA (Zhong et al., 2020), PIL-QA (Sovrano et al., 2021), JuRO (Craciun et al., 2025)), retrieved from a response bank (LegalQA (Askari et al., 2024), FALQU (Mansouri and Campos, 2023)) or generated from scratch. Examples of the latter include cLegal-QA (Wang et al., 2024c) and LeDQA (Liu et al., 2024) for Chinese, PrivacyQA (Ravichander et al., 2019) for English, LEGAL-UQA (Faisal and Yousaf, 2024) for Urdu, and GerLayQA (Büttner and Habernal, 2024) for German. These datasets often contribute to large legal benchmarks like LegalBench (Guha et al., 2023), which attempt to assess the legal knowledge and reasoning abilities of LLMs in a closed-book setting.

Legal RAG With the rapid rise of LLMs, RAG has emerged as a crucial method for improving the factual accuracy and interpretability of LQA systems (He et al., 2026). Combining the retrieval and generation steps, RAG aims to ground the response of an LLM in the retrieved context, making it a specific instance of open-book question answering (QA). Existing RAG resources usually consist of QA pairs (synthetic or human-written) labeled with relevant articles or passages from a large corpus of legal documents. Examples include LLeQA (Louis et al., 2024) for French, NitiBench (Akarajaradwong et al., 2025) for Vietnamese, and ObliQA (Gokhan et al., 2025) together with its multi-passage version ObliQA-MP (Gokhan and Briscoe, 2025) for English. For benchmarking, LegalBench-RAG (Pipitone and Alami, 2024) was proposed to assess precise retrieval by focusing on extracting minimal, highly relevant text segments from legal documents.

Our work builds on LLeQA (Louis et al., 2024), a long-form legal question-answering dataset in French that was primarily created by adding human-written responses to the BSARD (Louis and Spanakis, 2022) retrieval dataset. Similarly to bBSARD (Lotfi et al., 2025b) which extends BSARD to a bilingual French-Dutch setting, in this work we align, translate, and refine LLeQA annotations to create bLLeQA, and then use it to benchmark LLMs on LQA.

3 Dataset

In this section, we describe how bLLeQA was constructed from LLeQA (Louis et al., 2024). LLeQA comprises 1,868 expert-annotated legal questions in the French language, along with answers grounded in Belgian legislation (∼28k articles). The dataset was curated in collaboration with Droits Quotidiens[3], a Belgian non-profit organization that aims to make the law comprehensible and accessible to the public, and to this end, maintains a rich website featuring legal questions commonly posed by Belgian citizens. Each question comes with one or more categories, references to relevant legislative statutes, and a detailed answer written in layman's terms by experienced legal experts.

To create a parallel bilingual dataset from LLeQA, we follow these steps: (i) we extract and align the Dutch version of legal codes to build a parallel corpus; (ii) we leverage a combination of automatic translation and expert post-editing to translate questions and answers into Dutch; (iii) we ask a bilingual legal expert to ensure that answers are indeed grounded in the provided articles. Since LLeQA is based on BSARD, which has already been extended to a bilingual version (bBSARD (Lotfi et al., 2025b)), we take advantage of this resource where possible. These steps are described in more detail below.

Corpus Alignment The parallel bBSARD retrieval corpus covers 79% of the articles in LLeQA. For the rest, we scraped approximately 6,000 French-Dutch article pairs from the Justel Databank[4] to ensure that both language versions correspond to the same official legal provisions. These steps resulted in an alignment rate of 93% (25,982 out of 27,942) articles in both languages. None of the missing articles are cited in the answers.

QA Pairs Similarly to the previous step, we started with bBSARD, which covers 40% of the questions in LLeQA. Translations for the remaining questions and all reference answers were generated using `GPT-5.0`. Then a bilingual legal expert (native French and Dutch speaker) was asked to review the translations and rectify any potential issues[5].

[3] https://www.droitsquotidiens.be/fr

[4] https://www.ejustice.just.fgov.be

[5] In total, 20.8% of the reference answers and 1.4% of the questions needed corrections. Most common issues included failing to translate abbreviations (e.g. 'MENA') and to identify Flemish equivalents for Walloon institutions (e.g. 'FAMIWAL').

Annotations	Train	Val	Test	All
Initial dataset (#)	1,472	201	195	1868
No changes	70.3%	80.1%	70.3%	72.0%
Corrected	6.2%	10.0%	9.2%	6.3%
Removed	23.5%	9.9%	20.5%	21.7%
subject mismatch	2.1%	3.5%	8.7%	2.9%
general context	0.1%	0%	5.1%	0.6%
missing information	0.5%	0.5%	3.6%	0.8%
legal type mismatch	0.7%	0.5%	3.1%	0.9%
very long context	19.8%	5.5%	0%	16.2%
Final dataset (#)	1,125	181	155	1461

Table 1: RAG annotation outcomes across dataset splits.

RAG Annotations For the RAG setup, we ask the legal expert to check whether the cited articles provide sufficient context to produce an answer for each query. As Table 1 shows, 72% of the samples pass this step unchanged (accurate grounding). Over-citing samples (6.3%) were corrected by removing the unnecessary articles, and samples with more serious issues or an excessively long context were removed (21.7%). Most common issues include: (i) *subject mismatch*, where the cited articles do not address the legal subject of the question; (ii) *overly general context*, where the cited articles lack the specificity required to support the answer; (iii) *missing information*, where key information required to justify the answer is absent from the cited articles; (iv) *legal type mismatch*, where the cited articles concern procedural law while the question targets substantive provisions, or vice versa.

The final dataset comprises 1,461 QA pairs (for each language) grounded in a corpus of 25,982 articles of Belgian legislation. Appendix A contains additional details about the dataset.

4 Experimental Setup

This section describes the experimental setup for benchmarking retrieval, reranking, and end-to-end RAG on bLLeQA. All experiments are conducted in a zero-shot setting on the test set.

4.1 Retrieval and Reranking

Retrieval Retrieval experiments are based on the code[6] from bBSARD (Lotfi et al., 2025b). We select a wide range of models, from lexical approaches to static embeddings (i.e., `word2vec` and `fastText`), and zero-shot dense retrievers. The complete list of models and the prompts used for the instruct models are provided in Tables 13 and

[6] https://github.com/nerses28/bBSARD

15 (Appendix D and E), respectively. For models with a maximum input length of 512 tokens, texts are split into overlapping chunks of 200 tokens with a 20-token overlap. Embeddings from each chunk are aggregated using mean pooling (except for `LaBSE` which uses the `[CLS]` token). Then, cosine similarity is computed to score the resulting embeddings.

Reranking In addition, we benchmark a number of reranking models (see Table 13 in Appendix D) using the top 100 articles retrieved by `BM25`, the `E5` suite, and `Voyage`.

Evaluation Metrics The models were evaluated using conventional retrieval metrics: macro-averaged recall@k (R@k), mean average precision@k (MAP@k), mean reciprocal rank@k (MRR@k), and normalized discounted cumulative gain@k (nDCG@k).

4.2 Question Answering

Generating Answers We prompt a wide selection of LLMs, both open-weight and proprietary (see Table 13 in Appendix D), to generate answers to Dutch and French legal questions. For each question, we experiment with three different context settings:

1. *Gold*: A setting where we provide only the *gold-standard context*. We use this setting as a baseline to validate whether the models can follow the instructions and use correct articles.
2. *RAG*: We conduct retrieval on the Dutch corpus with `voyage-3-large` and use the retrieved *top 100* article IDs for both French and Dutch. The context contains all relevant articles in 71% of cases.
3. *RAG+*: Since the *top 100* retrieved articles do not always include the full gold context (Recall@100 < 1), we add any missing gold-standard articles to the context by replacing an equal number of randomly selected retrieved articles.

The prompt for the models, which remains consistent across settings and models, can be found in Table 16, Appendix E. We do not apply any truncation policy, as the full context in all settings fits within the context window of the tested LLMs. Importantly, we instruct models to (i) refuse to answer a question when the context is insufficient and (ii) answer in paragraphs supported by one or more articles from the provided context. We use the OpenRouter API[7] to query the models.

Evaluation Metrics We evaluate the output of the models on four different aspects.

Correctness: We leverage DeepEval's G-Eval metric (Yang et al., 2024; Liu et al., 2023) to evaluate answer correctness, using an LLM-as-judge with a custom prompt (cf. Table 14, Appendix E). In essence, the LLM is tasked with assigning a score to model outputs ranging from 1 to 5 (1: Critical Failure/Incorrect, 2: Poor/Significant Omissions, 3: Acceptable/Partially Complete, 4: Good/Mostly Accurate, 5: Excellent/Semantically Equivalent), given the gold standard answer. We report averages across all queries per model.

Faithfulness: We evaluate the outputs using RAGAS' faithfulness metric (Es et al., 2024). This metric leverages an LLM-as-judge to extract statements from a model's answer, and verify the proportion of statements that are supported by the context to produce a score in $[0, 1]$. We compute the metric only with respect to the articles cited by the tested models.

Citation: Since models are prompted to explicitly ground their statements in context articles, we can compare the cited articles with the gold set to calculate precision, recall, and F1 scores. These metrics provide insight into how well models are able to extract the relevant information from the provided context.

Refusal: We also prompt the models to answer the question only if the context is sufficient, and refuse otherwise. This can be used to assess the models' ability in determining whether the provided context is adequate to answer a question, which is a desirable feature, especially when dealing with sensitive and potentially consequential domains like law. Comparing the true incomplete (or inadequate) contexts with the refused responses, we calculate and report *Precision*, *Recall*, and *F1* scores.

In all settings, correct and incorrect refusal are automatically scored with the upper and lower bound of the correctness and citation scales, respectively: a correct refusal receives the maximum scores for citation coverage (1) and correctness (5), whereas an incorrect refusal receives the minimum scores of 0 for citation coverage and 1 for correctness.

[7] `https://openrouter.ai/`

Model	Size	French			Dutch		
		R@100	MAP@100	NDCG@100	R@100	MAP@100	NDCG@100
BM25	-	57.04	19.25	28.44	48.52	15.46	22.93
BM25 + BGE-reranker-v2-m3	-	57.04	27.01	35.24	48.52	22.92	29.93
mE5-small	118M	58.44	19.09	28.10	60.64	21.87	30.85
mE5-base	278M	58.94	20.71	29.77	61.53	21.03	30.48
mE5-large	560M	66.84	27.80	37.31	68.01	25.64	35.94
mE5-large-instruct	560M	73.46	22.49	34.30	71.37	26.02	36.97
E5-mistral-7b	7B	74.30	**36.91**	46.45	74.73	**36.37**	**46.27**
BGE-Mult.-Gemma2	9B	79.86	36.22	**48.01**	79.59	31.25	43.57
voyage-3-large	-	**81.84**	34.16	46.51	76.55	32.46	43.65
embedding-3-large	-	79.05	33.61	45.33	**80.94**	27.54	40.66

Table 2: Retrieval performance of selected models.

Selecting a Judge Model Choosing an appropriate judge model is quintessential for accurately estimating a model's RAG capabilities. To ensure that the model's assessments align with human judgment, we use all responses produced by `DeepSeek-v3.2` in the *Gold* setting[8], and asked a legal expert to score them from 1 to 5, following the instructions in the answer correctness prompt (Table 14, Appendix E). We then prompt a list of candidate LLMs[9] in the same way, and calculate Spearman correlation, mean average error (MAE) and F1-macro (for the binarized correctness scale), with respect to the expert annotations. In our experiment, `Gemini-3-Flash` achieves a strong correlation, highest F1-macro, and lowest MAE, for both French and Dutch, and therefore is chosen as the judge. The results for all candidates can be found in Table 5, Appendix B.

5 Results and Discussion

In this section, we present and discuss the main experimental results for legal RAG, in both retrieval and generation. In particular, we examine differences in performance between proprietary and open-weight models.

5.1 Retrieval

Table 2 shows the retrieval results for the selected models (detailed results are provided in Tables 6 and 7 in Appendix C). `BM25` achieves performance comparable to multilingual `E5-small` in French, but performs worse than it in Dutch. When combined with a reranker, `BM25` outperforms multilingual `E5-base` by a large margin in French and performs comparably in Dutch (MAP@100 and NDCG@100). In general, model performance tends to improve with model size in both languages. Interestingly, large open-weight models perform comparably to, and in some cases better than proprietary models in both languages.

5.2 Generation

In this section, we discuss the generative performance of LLMs in terms of the correctness and faithfulness of the response, citation recall, and refusal, as described in Section 4.2.

Robustness to Noise Figure 1 visualizes the quality of answer generation per model and setup. To better understand the results, we divide the responses into 4 categories: correct and incorrect refusals, and accurate and inaccurate answers. To be considered accurate, an answer should score 4 or higher on the 1-5 correctness scale (see the evaluation prompt in Table 14, Appendix E).

Surprisingly, the *Gold* setting does not yield the highest proportion of accurate answers across all models. In many cases, when noise (i.e. irrelevant articles) is introduced into the context in RAG+, the proportion of accurate answers increases (e.g. `Qwen3.5-27B`, `GLM-5`, `Claude-Sonnet-4.6`). This effect might be attributed to lower refusal rates when a longer context is provided. This can suggest a context size bias, where a longer context has a higher chance to pass as sufficient, even if it contains the same necessary information as a significantly shorter context.

In the realistic RAG setting, we observe a slight decrease in the proportion of accurate answers for many models compared with the RAG+ setup (e.g. `GLM-5`, `GPT-5.4`). However, for models with relatively strong refusal capabilities, such as `GLM-5` and `GPT-5.4`, the combined proportion of accurate answers and correct refusals can match or exceed the accurate answer rate observed in the earlier setups, where correct refusals do not apply. Hence,

[8] We chose this setting to isolate the answer quality from retrieval errors.

[9] We select recent, high-performing LLMs, as well as models identified in recent studies on LLM-as-a-judge, including Han et al. (2026) and Feng et al. (2025).

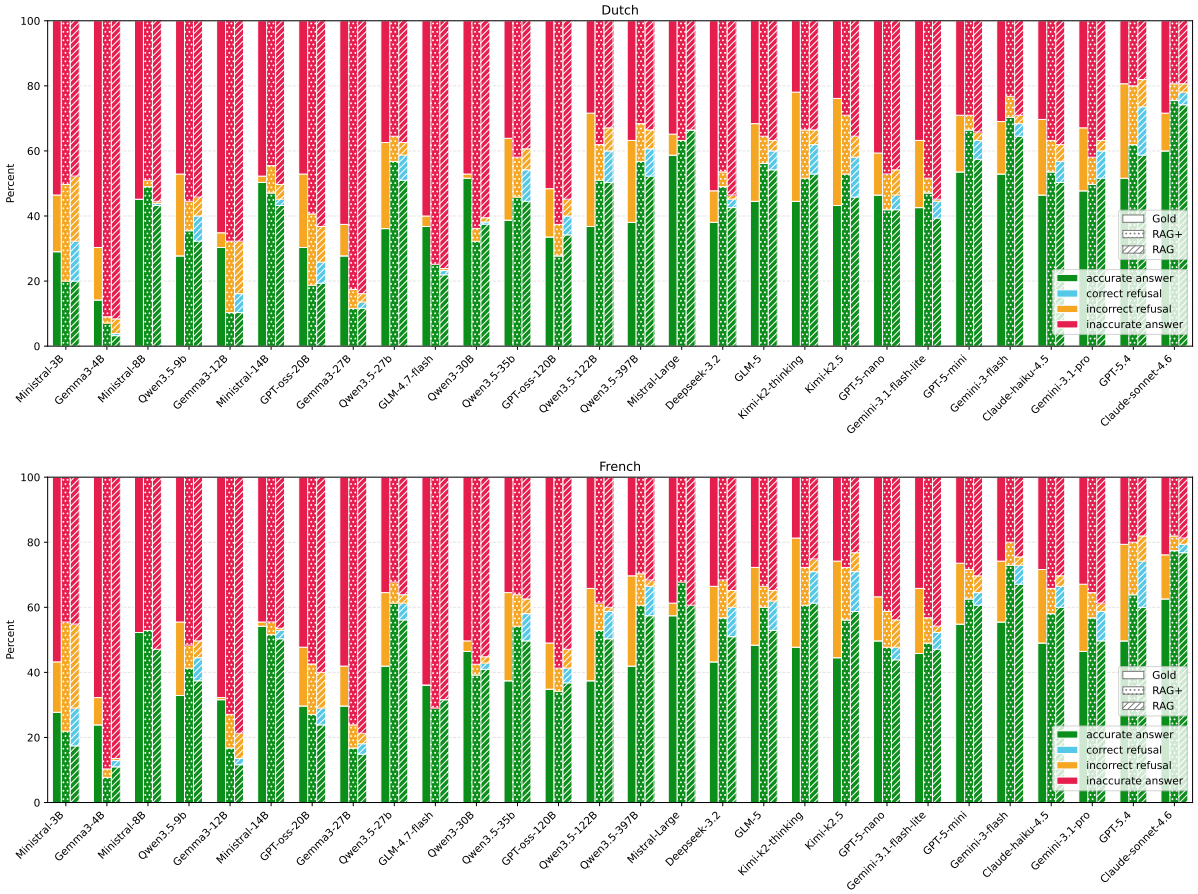

Figure 1: Performance of models on the test set under different context settings for Dutch (top) and French (bottom). Table 10 in Appendix C contains the exact results that correspond to this figure.

we can conclude that modern state-of-the-art models of different sizes are generally robust to noise and can perform comparably across settings with larger contexts.

Citation Figure 2 shows the average answer correctness score against citation recall under the three context settings, for Dutch and French. As can be observed, there is a strong positive correlation between the two, especially under the RAG and RAG+ settings. While in general this trend adheres to the performance-size relation (i.e. larger models doing better), there are exceptions, most notably the `Qwen3.5` family, which performs surprisingly well for its size. In particular, the 27B version achieves relatively high recall scores in a realistic RAG setting, comparable to the largest open models (e.g. `GLM-5`) and even some state-of-the-art proprietary models (e.g. `Gemini-3.1-Pro`). This suggests a cost-effective strategy for RAG design by leveraging their citation selection capability, while delegating the generation part to a larger model.

Generation Quality Figure 3 plots the average answer correctness score against the corresponding average faithfulness score, under the three context settings, for Dutch and French. Overall, higher faithfulness is associated with higher correctness, but the relationship shows substantial local variation across models and settings. Consequently, better overall performance is indicated by moving toward the upper-right corner (high faithfulness and high correctness).

For French and Dutch, we observe that proprietary models (`GPT-5.4`, `Claude-Sonnet-4.6`, `Gemini-3-Flash`) achieve the highest performance in the RAG setup, combining strong faithfulness and correctness scores. The best observed mean correctness is 4.2, corresponding to the "Good/Mostly Accurate" category. At the same time, roughly 20% of the outputs are flawed, i.e., inaccurate answers or incorrect refusals. The best open-weight models (`GLM-5`, `Qwen3.5-397B-A17B`, `Kimi-K2-Thinking`) achieve faithfulness scores comparable to the strongest proprietary models, but

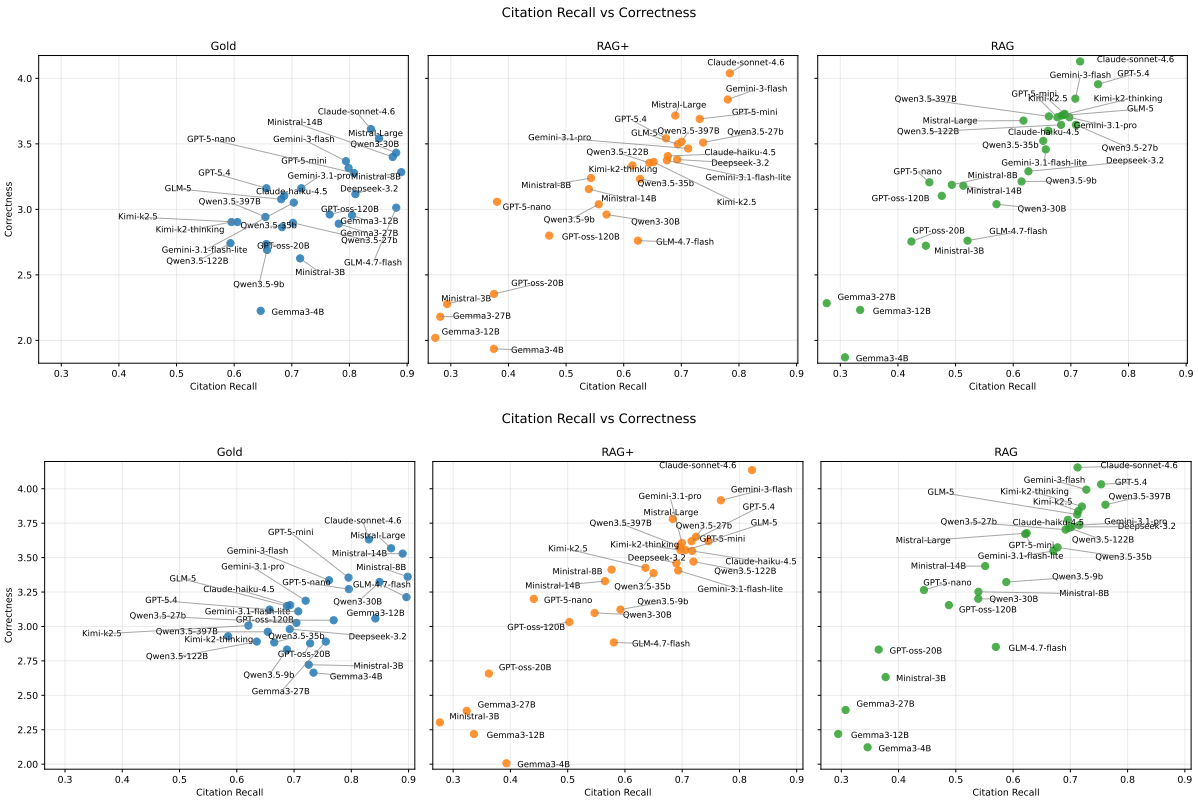

Figure 2: Answer correctness (Table 9, Appendix C) vs. citation recall (Table 8, Appendix C) for different models under the 3 context settings, for Dutch (top) and French (bottom).

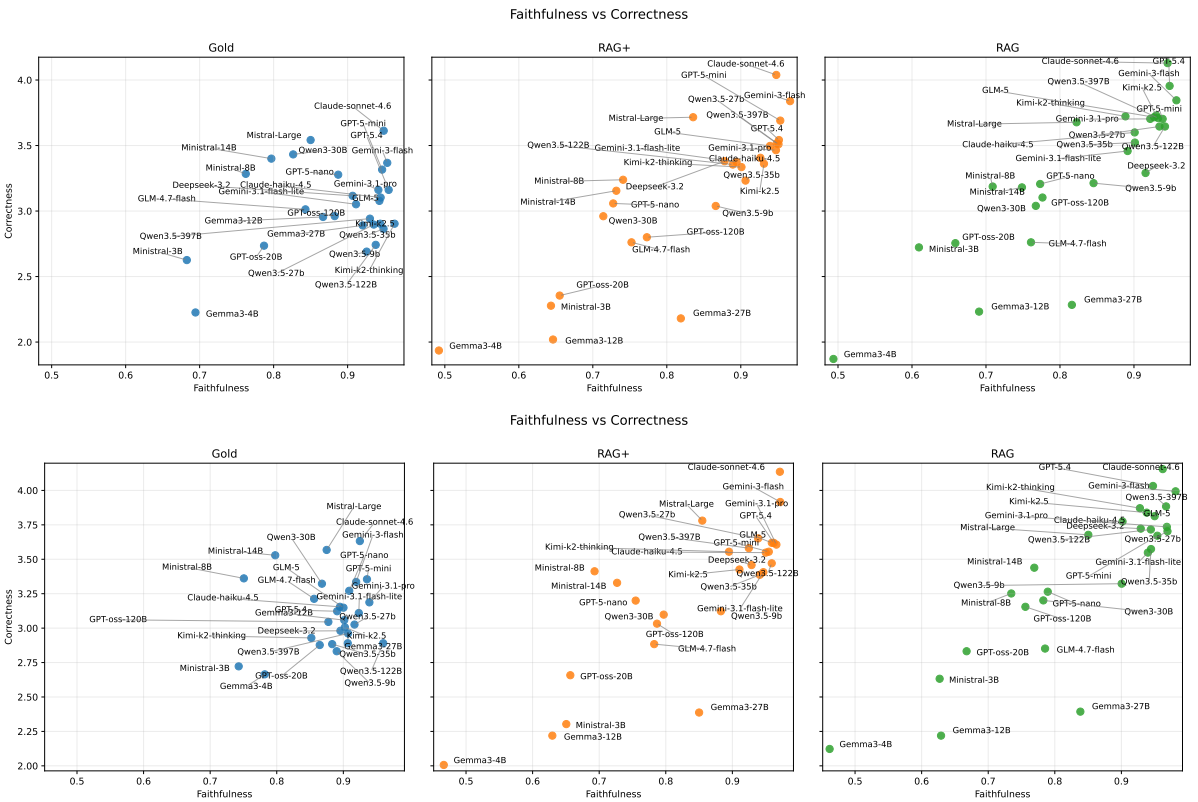

Figure 3: Answer correctness vs. answer faithfulness (Table 9, Appendix C) for different models under the 3 context settings, for Dutch (top) and French (bottom).

this does not translate into the same level of answer quality: their average correctness remains below 4 (the "Good/Mostly Accurate" threshold). Their best-case correctness scores are 3.9 for French and 3.7 for Dutch. For these models, the share of flawed outputs is around 35%. Hence, we observe a large gap between the best proprietary and open-weight models in the real-world setup for both Dutch and French. At the same time, the best open-weight models match or outperform smaller proprietary models such as `GPT-5-mini` and `Claude-Haiku-4.5`.

For both languages, we observe strong results for `Qwen3.5`-family models in all settings. Starting from the 27B version, the `Qwen3.5`-family combines high faithfulness with strong correctness of 3.7 for French and Dutch, placing the results close to the "Good/Mostly Accurate" category.

Refusal We observe that all models struggle to detect incomplete contexts, resulting in poor overall refusal performance (see Tables 11 and 12, Appendix C). The best proprietary (`GPT-5.4`) and open (`Kimi-K2.5`) models reach RAG refusal F1 scores of around 56 and 52, respectively. Other models demonstrate even less promising performances. This behavior in refusal handling or context completeness detection is not unique to our case and has also been observed in other tasks (Xu et al., 2025; Sun et al., 2025; Zhou et al., 2026; Kirichenko et al., 2025). Notably, poor refusal performance does not necessarily imply low correctness, as two of the three best-performing models in the RAG setup, `Claude-Sonnet-4.6` and `Gemini-3-Flash`, exhibit low refusal performance. The models' strong performance in faithfulness indicates that they ground their responses in the provided articles rather than drawing on parametric memory. This suggests that the missing ground-truth articles in the context are not equally important, and in some cases accurate answers can be generated using an incomplete context. Despite this, such behavior can be problematic, as models may overlook important details and nuances that are particularly important in the legal domain.

6 Conclusions and Future Work

In this paper, we presented bLLeQA, a parallel bilingual dataset for retrieval-based LQA in French and Dutch. Based on the LLeQA dataset, it comprises 1461 QA pairs (for each language) grounded in a corpus of 25,982 articles of Belgian legislation, providing a valuable resource for LQA studies in French and Dutch. Using bLLeQA, we evaluated a wide range of open and proprietary LLMs on legal question-answering in both languages, under three different RAG settings, assessing their ability to generate correct and grounded responses, or to refuse when the provided context is insufficient. Open-weight models are competitive with proprietary models in retrieval and citation extraction but still lag behind the strongest proprietary models in response generation. Across all models, correct refusal capability remains weak: models do not reliably detect when the provided sources are incomplete, and this failure leads to performance degradation and an increased proportion of erroneous answers. Moreover, even in the best-case scenario, the end-to-end RAG setup yields a substantial proportion of flawed responses (20%). We believe these findings provide guidance for designing RAG-based LQA systems in French and Dutch.

There are many avenues worth further exploration. Most interesting to us is a detailed error analysis of hallucinations in LLMs: while partly captured by *faithfulness* and *correctness*, it warrants more targeted study for a finer-grained assessment. In addition, we did not explore fine-tuning or agentic setups. Both directions could help address the failure modes observed in our naive RAG pipeline, for example by improving citation extraction and refusal reliability, or by enabling retrieval and verification steps to reduce unsupported generations under noisy or incomplete context.

Limitations

bLLeQA offers limited coverage of Belgian law, focusing on selected codes from federal and Walloon legislation. In addition, it reflects a specific time slice corresponding to when the original LLeQA dataset was constructed. Given these limitations, bLLeQA is not intended to provide comprehensive legal information or advice. Instead, its primary purpose is to benchmark retrieval and generative models grounded in the provided sources and to gain insights into the current state of the art.

Acknowledgements

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. We thank legal professional Manon Quinet for her assistance with data annotation. In addition, we

acknowledge the use of ChatGPT for assisting with error checking and proofreading of this paper.

References

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34304–34327, Suzhou, China. Association for Computational Linguistics.

Anthropic. 2025. Claude haiku 4.5. https://www.anthropic.com/claude/haiku. Accessed: 2026-02-09.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-03-27.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Comput. Surv.*, 58(6).

Arian Askari, Zihui Yang, Zhaochun Ren, and Suzan Verberne. 2024. Answer retrieval in legal community question answering. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III*, page 477–485, Berlin, Heidelberg. Springer-Verlag.

{Nigel J.} Balmer, Ash Patel, Alexy Buck, Catrina Denvir, and Pascoe Pleasence. 2010. *Knowledge, Capacity and the Experience of Rights Problems*. Public Legal Education Network: PLENet.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027, St. Julian's, Malta. Association for Computational Linguistics.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 71–80, New York, NY, USA. Association for Computing Machinery.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. 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.

Cristian-George Craciun, Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, and Mihaela-Claudia Cercel. 2025. GRAF: Graph retrieval augmented by facts for Romanian legal multi-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12708–12742, Vienna, Austria. Association for Computational Linguistics.

DeepMind. 2025a. Gemini 2.5 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Lite-Model-Card.pdf. Accessed: 2026-03-27.

DeepMind. 2025b. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Accessed: 2026-02-10.

DeepMind. 2025c. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Accessed: 2026-02-10.

DeepMind. 2026. Gemini 3.1 flash-lite model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Lite-Model-Card.pdf. Accessed: 2026-03-27.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, Heng Wang, and Zhiyuan Liu. 2019. *CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension*, page 439–451. Springer International Publishing.

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.

Faizan Faisal and Umair Yousaf. 2024. Legal-uqa: A low-resource urdu-english dataset for legal question answering. *Preprint*, arXiv:2410.13013.

Jean-Philippe Fauconnier. 2015. French word embeddings. Accessed: 2026-01-05.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Yuanning Feng, Sinan Wang, Zhengxiang Cheng, Yao Wan, and Dongping Chen. 2025. Are we on the right way to assessing llm-as-a-judge? *Preprint*, arXiv:2512.16041.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Tuba Gokhan and Ted Briscoe. 2025. Grounded answers from multi-passage regulations: Learning-to-rank for regulatory RAG. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 135–146, Suzhou, China. Association for Computational Linguistics.

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2025. Shared task RIRAG-2025: Regulatory information retrieval and answer generation. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 1–4, Abu Dhabi, UAE. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. 2026. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement.

Congqing He, Haichuan Hu, Yanli Li, Hao Zhang, and Quanjun Zhang. 2026. A survey of large language models for legal tasks: Progress, prospects and challenges. *Computer Science Review*, 60:100906.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913, Albuquerque, New Mexico. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

JinaAI. 2025. jinaai/jina-reranker-v2-base-multilingual. `https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual`. Accessed: 2026-02-09.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *Preprint*, arXiv:2312.15503.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2251–2260, New York, NY, USA. Association for Computing Machinery.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. Ledqa: A chinese legal case document-based question answering dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5385–5389, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ehsan Lotfi, Nikolay Banar, and Walter Daelemans. 2025a. BEIR-NL: Zero-shot information retrieval benchmark for the Dutch language. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 36–45, Abu Dhabi, UAE. Association for Computational Linguistics.

Ehsan Lotfi, Nikolay Banar, Nerses Yuzbashyan, and Walter Daelemans. 2025b. Bilingual BSARD: Extending statutory article retrieval to Dutch. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 10–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Antoine Louis, Vageesh Kumar Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2025. ColBERT-XM: A modular multi-vector representation model for zero-shot multilingual information retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4370–4383, Abu Dhabi, UAE. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22266–22275.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2342–2348, New York, NY, USA. Association for Computing Machinery.

Behrooz Mansouri and Ricardo Campos. 2023. Falqu: Finding answers to legal questions. *Preprint*, arXiv:2304.05611.

Meta. 2025. Llama-3.3-70b-instruct model card. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2026-02-10.

Meta. 2026. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-03-27.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mistral AI. 2025. Mistral large 3-675b instruct-2512 model card. https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512. Accessed: 2026-02-10.

Moonshot AI. 2026. Kimi k2.5: Ai that sees, codes, and works like an expert. https://www.kimi.com/ai-models/kimi-k2-5. Accessed: 2026-03-27.

OpenAI. 2025. Gpt-5 mini model. https://platform.openai.com/docs/models/gpt-5-mini. Accessed: 2026-02-10.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025b. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-11-13.

OpenAI. 2025c. Introducing gpt-5.4. https://openai.com/index/us-EN/introducing-gpt-5.4/. Accessed: 2026-03-27.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

QwenTeam. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-03-27.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. 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 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Felicia Redelaar, Romy Van Drie, Suzan Verberne, and Maaike De Boer. 2024. Attributed question answering for preconditions in the Dutch law. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 154–165, Miami, FL, 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.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. pages 0–.

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 230–234, New York, NY, USA. Association for Computing Machinery.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2025. Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, page 123–129, Berlin, Heidelberg. Springer-Verlag.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10658–10671, Miami, Florida, USA. Association for Computational Linguistics.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11461–11480.

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, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025b. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4130–4136, Portorož, Slovenia. European Language Resources Association (ELRA).

Stephan Tulkens and Thomas van Dongen. 2024. Model2vec: Fast state-of-the-art static embeddings.

Santosh T.y.s.s., Rashid Haddad, and Matthias Grabmair. 2024. ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the European court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.

VoyageAI. 2024. Domain-specific embeddings and retrieval: Legal edition (voyage-law-2). VoyageAI blog. Accessed: 2025-11-13.

VoyageAI. 2025. Voyage 3 large. `https://blog.voyageai.com/2025/01/07/voyage-3-large/`. Accessed: 2025-11-13.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. Technical Report MSR-TR-2024-45, Microsoft.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Yizhen Wang, Xueying Shen, Zixian Huang, Lihui Niu, and Shiyan Ou. 2024c. clegal-qa: a chinese legal question answering with natural language generation methods. *Complex & Intelligent Systems*, 11.

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, page 641–649, New York, NY, USA. Association for Computing Machinery.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025. Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9541–9564.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can large multimodal models uncover deep semantics behind images? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1898–1912, Bangkok, Thailand. Association for Computational Linguistics.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *Preprint*, arXiv:2412.04506.

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

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024b. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9701–9708.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. 2026. Do retrieval augmented language models know when they don't know? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35158–35166.

A bLLeQA

Table 3 shows the topic distribution of questions in bLLeQA. Table 4 and Figure 4 show the distribution of articles for the codes that contain relevant instances in the dataset. Figure 5 presents key statistics for the bLLeQA dataset.

Topic	Train	Val	Test	(%)
Housing	327	49	74	30.8
Healthcare	191	37	48	18.9
Family	175	21	14	14.4
Work	114	24	7	9.9
Immigration	122	18	1	9.6
Money	99	13	4	7.9
Privacy	69	12	7	6.0
Justice	28	7	0	2.3

Table 3: Topic distribution of questions in bLLeQA.

B LLM as a Judge

Table 5 reports the agreement between model predictions and human judgments.

C Additional Results

Tables 6, 7, 8, 9, 10, 11 and 12 show the detailed results of the retrieval, reranking and RAG experiments.

D Models

Table 13 presents the models we used in our experiments, as well as their sizes and citations.

E Prompts

Tables 14, 15 and 16 show the prompt templates used for retrieval, generation and evaluation.

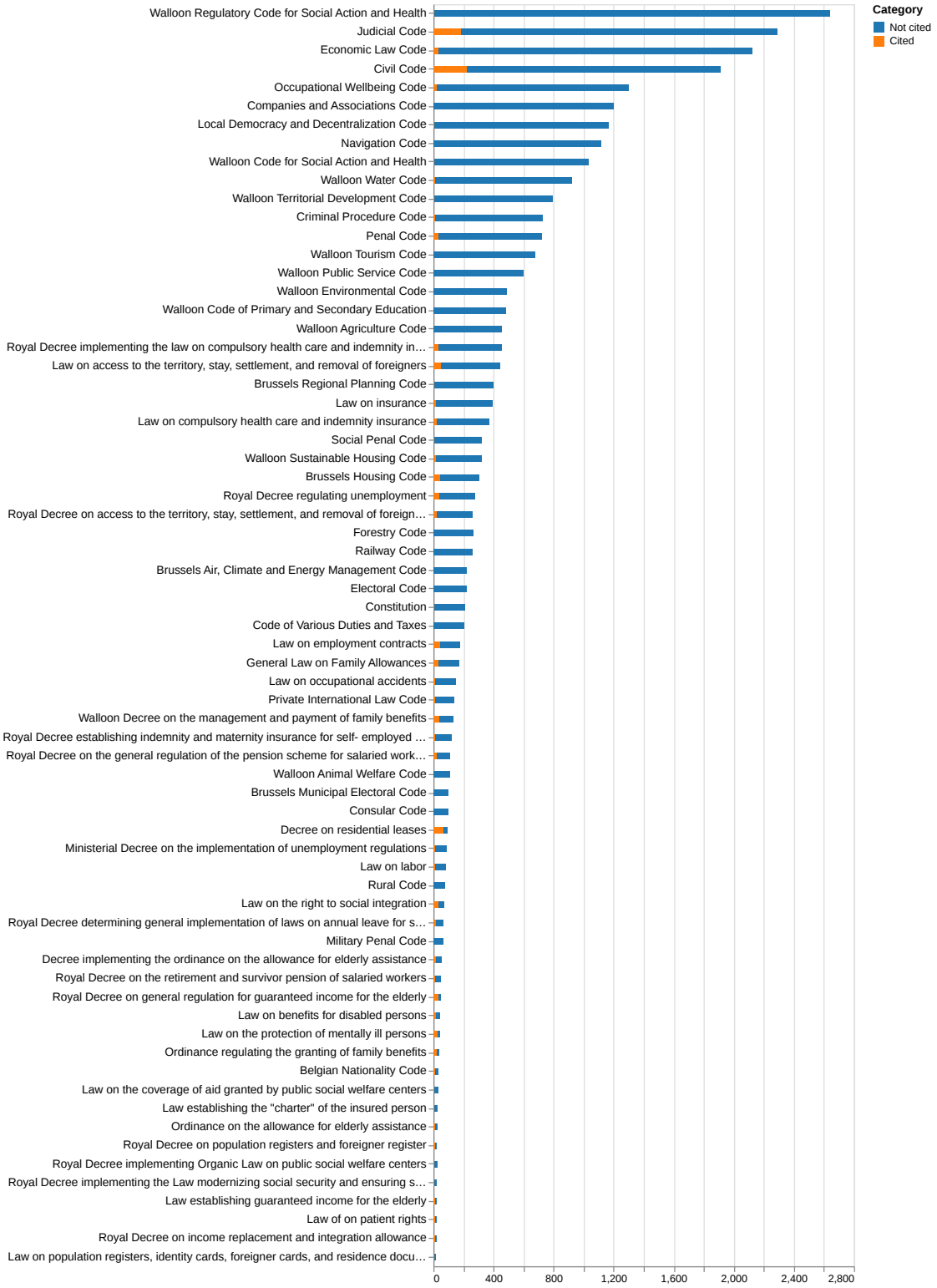

Figure 4: Distribution of codes in the bLLeQA corpus. Articles labeled as “Cited” appear in the training, validation, and test sets, whereas “Not cited” articles do not. Code names are translated for readability.

Code	Total	Relevant
Walloon Regulatory Code for Social Action and Health	2462	1
Judicial Code	2017	156
Economic Law Code	1921	24
Civil Code	1719	158
Code on well-being at work	1270	22
Code of Companies and Associations	1123	0
Code of Local Democracy and Decentralization	1110	3
Walloon Social Action and Health Code	978	3
Belgian Navigation Code	958	0
Environment Code – Water Code – Decree Section	867	7
Walloon Code of Territorial Development	786	4
Penal Code	680	26
Criminal Procedure Code	674	4
Walloon Tourism Code	600	0
Walloon Civil Service Code	490	0
Royal Decree implementing the law on compulsory health insurance and benefits, coordinated on 14 July 1994	448	30
Walloon Agriculture Code	447	0
Law of 15 December 1980 on entry into the territory, residence, establishment, and removal of foreign nationals	440	47
Brussels Code on Spatial Planning	395	1
Insurance Act	390	14
Law on compulsory health insurance and benefits, coordinated on 14 July 1994	369	65
Walloon Code on Sustainable Housing	305	16
Social Penal Code	299	3
Royal Decree regulating unemployment	275	37
Brussels Housing Code	260	36
Railway Code	260	0
Forest Code	259	0
Royal Decree on the entry, stay, settlement and removal of foreign nationals	252	21
Walloon Code on Primary and Secondary Education	238	0
Electoral Code	219	0
Brussels Code on Air, Climate and Energy Management	207	0
The Constitution	206	1
Walloon Environmental Code	198	0
Law on employment contracts	173	40
Environmental Code	172	0
General Law on Family Allowances	168	32
Codes on Miscellaneous Rights and Taxes	162	0
Code on Primary and Secondary Education	151	0
Law on Work Accidents	148	11
Code of Private International Law	132	6
Decree on the management and payment of family benefits	131	35
Royal Decree establishing compensation insurance and maternity insurance for self-employed workers and assisting spouses	120	9
Royal Decree establishing general regulations for the retirement and survivor's pension scheme for salaried workers	110	19
Walloon Animal Welfare Code	106	0
Consular Code	100	0
Brussels Municipal Electoral Code	98	0
Ministerial decree laying down the procedures for implementing unemployment regulations	88	9
Labor Law	85	9
Rural Code	75	0
Royal Decree determining the general terms and conditions for the implementation of laws relating to annual leave for salaried workers	68	13
Law concerning the right to social integration	68	29
Law containing the Military Penal Code	65	0
Decree relating to residential leases	60	44
Royal Decree No. 50 on the Old-Age and Survivors' Pensions for Salaried Workers	52	9
Royal Decree establishing general regulations on income guarantees for elderly persons	49	6
Decree of the Joint Community Commission implementing the Order of 10 December 2020 on allowances for assistance to elderly persons	44	14
Law on allowances for persons with disabilities	43	15
Law on the protection of persons with mental disorders	42	23
Ordinance regulating the granting of family benefits	40	20
Law on the provision of assistance by public social welfare centers	31	5
Belgian Nationality Code	29	1
Environment Code - Book 2: Water Code. - Decree section	27	1
Law establishing the Charter of the Socially Insured Person	25	5
Ordinance on allowances for assistance to the elderly	25	8
Royal Decree of 16 July 1992 on population registers and the register of foreign nationals	24	9
Royal Decree of 9 May 1984 implementing Article 100bis, §1, of the Organic Law of 8 July 1976 on public social assistance centers	24	4
Royal Decree implementing Articles 15, 16 and 17 of the Law of 26 July 1996 on the modernization of social security and ensuring the viability of statutory pension schemes	21	5
Decree of 15 March 2018 on residential leases	19	19
Law establishing income support for the elderly	19	8
Royal Decree on income replacement allowance and integration allowance	18	11
Law on patient rights	15	5
Law of 19 July 1991 on population registers, identity cards, foreigner cards and residence permits	15	3
Decree of the Joint Community Commission of 28 January 2021 implementing the Order of 10 December 2020 on the allowance for assistance to the elderly	10	0
Law of 22 August 2002 on patients' rights	3	2
Royal Decree of 25 November 1991 regulating unemployment	2	0
Royal Decree of 21 December 1967 laying down general regulations for the retirement and survivor's pension scheme for salaried workers	2	2
Decree of the Walloon Region of 8 February 2018 on management and payment	1	1

Table 4: Distribution of legal articles by code. "Total" denotes the number of occurrences in the full corpus, and "Relevant" denotes the number of occurrences among articles referenced across the train/validation/test splits. Code names are translated for readability.

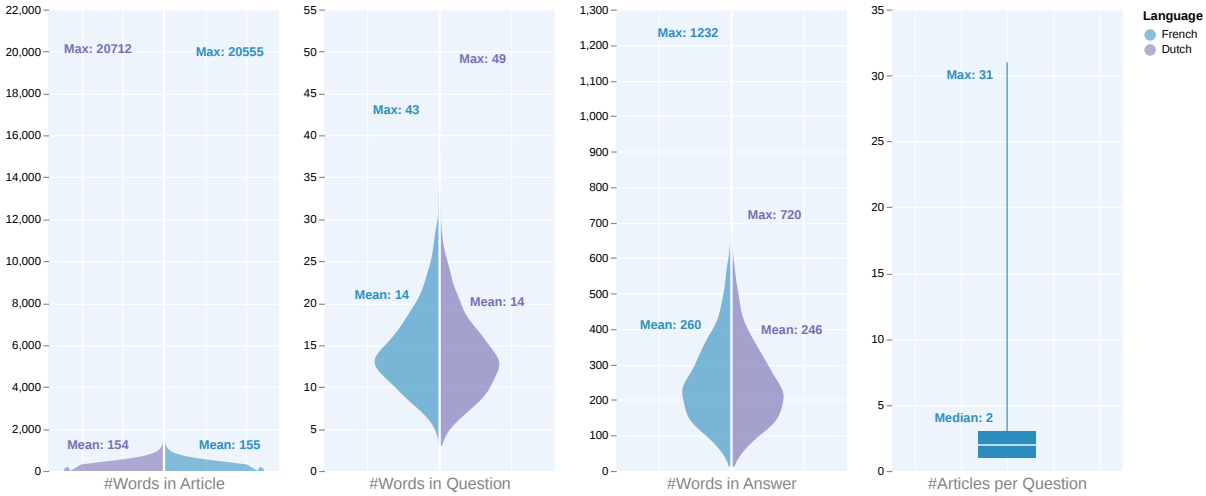

Figure 5: Basic statistics of bLLeQA. From the left: number of words in the articles (French and Dutch), number of words in the questions (French and Dutch), number of words in the answers (French and Dutch), and number of relevant articles per question.

Model	Size	French			Dutch		
		Spearman↑	MAE↓	F1-macro↑	Spearman↑	MAE↓	F1-macro↑
Gemma-3-4B-it	4B	0.49	1.39	49.97	0.30	1.10	55.57
Qwen3.5-9B	9B	0.77	1.30	41.17	0.58	0.93	53.69
Gemma-3-12B-it	12B	0.74	0.95	70.64	0.50	0.92	59.52
GPT-oss-20B	20B	0.71	1.66	37.02	0.46	1.36	49.98
Qwen3.5-27B	27B	0.78	1.15	51.72	**0.62**	0.89	57.51
Gemma-3-27B-it	27B	0.77	1.31	37.02	0.50	1.01	50.42
GLM-4.7-Flash	30B	0.69	1.51	38.71	0.43	1.25	41.45
Qwen3-30B-A3B-Instruct	30B	0.72	1.61	29.82	0.44	1.46	38.58
Qwen3.5-35B-A3B	35B	0.75	1.42	35.29	0.56	1.03	49.02
Llama-3.3-70B-Instruct	70B	0.71	1.64	36.16	0.54	1.34	51.86
Qwen3-Next-80B-A3B-Instruct	80B	0.67	1.62	37.02	0.44	1.45	46.06
Llama-4-Scout	109B	0.76	1.04	63.74	0.54	0.93	64.63
GPT-oss-120B	120B	0.74	1.27	55.87	0.55	1.02	59.43
Qwen3.5-122B-A10B	122B	0.81	1.37	37.87	0.59	1.00	48.98
Qwen3-235B-A22B	235B	0.73	1.55	27.90	0.49	1.27	42.96
Qwen3.5-397B-A17B	397B	0.76	1.39	37.02	0.55	1.08	49.48
Llama-4-Maverick	400B	0.74	1.07	59.87	0.53	0.89	62.98
Mistral-Large-2512	675B	0.76	1.21	46.62	0.54	0.92	53.53
DeepSeek-v3.2	685B	0.74	1.53	30.77	0.50	1.20	39.29
GLM-5	754B	0.80	1.36	41.17	0.54	1.05	49.98
Kimi-K2-Instruct-0905	1T	0.76	1.39	32.61	0.48	1.22	40.38
Kimi-K2-Thinking	1T	0.69	1.66	26.92	0.46	1.50	33.89
Kimi-K2.5	1.1T	0.79	1.41	31.67	0.56	1.03	41.89
Gemini-3.1-Flash-Lite	–	0.76	1.34	43.55	**0.62**	1.03	49.98
Claude-Haiku-4.5	–	0.74	1.63	29.82	0.50	1.40	37.04
Gemini-2.5-Flash	–	0.76	1.28	45.86	0.58	1.01	53.69
Gemini-3-Flash	–	0.74	**0.83**	**71.26**	0.60	**0.71**	**72.52**
Gemini-3.1-Pro	–	0.78	1.30	39.54	0.60	0.97	48.53
Claude-Sonnet-4.6	–	**0.82**	1.28	37.02	0.59	0.95	48.05

Table 5: Comparison of model judgments against human judgments on DeepSeek's output for French and Dutch. We report Spearman correlation, MAE, and F1-macro for accurate (scores of 4–5) versus non-accurate (scores of 1–3) answers. Models are ordered by size when available. G-Eval does not support OpenAI models, and Ministral models produced outputs in the wrong format.

Lang	Model	Size	R@100	R@200	R@500	MAP@100	MRR@100	NDCG@10	NDCG@100
	TF-IDF	-	59.24	63.11	72.63	18.68	20.73	22.31	27.82
	BM25	-	57.04	64.08	72.50	19.25	22.65	23.94	28.44
	word2vec	-	57.93	66.77	74.56	12.93	16.11	16.42	22.94
	fastText	-	29.78	34.40	45.72	6.70	8.59	9.41	11.99
	static-similarity-mrl-multilingual-v1	-	47.28	55.67	64.89	11.20	13.90	13.69	19.10
	mE5-small	118M	58.44	63.79	71.29	19.09	20.57	23.12	28.10
	potion-multilingual-128M	128M	44.00	55.88	73.56	7.97	9.80	9.21	15.47
	mContriever	178M	54.81	60.80	71.61	9.38	11.72	12.56	19.46
	DPR-XM	277M	38.70	46.98	56.47	8.41	11.29	11.37	15.34
	mE5-base	278M	58.94	64.43	71.95	20.71	22.91	25.37	29.77
	mGTE	305M	65.47	70.74	77.61	20.57	23.17	23.64	30.59
FR	LaBSE	471M	24.97	33.41	47.90	2.18	3.38	2.52	6.82
	mE5-large	560M	66.84	71.84	77.54	27.80	30.61	32.62	37.31
	mE5-large-instruct	560M	73.46	78.11	84.45	22.49	26.40	26.52	34.30
	BGE-M3	568M	67.30	73.00	78.19	20.45	23.02	25.50	31.58
	snowflake-arctic-embed-l-v2.0	568M	60.61	70.07	81.87	14.62	18.87	17.89	25.18
	jina-embeddings-v3	572M	63.60	71.66	79.61	16.86	20.70	18.80	27.09
	E5-mistral-7b	7B	74.30	77.91	80.70	**36.91**	41.07	42.61	46.45
	BGE-Mult.-Gemma2	9B	79.86	84.34	88.21	36.22	**43.35**	**42.88**	**48.01**
	voyage-2-law	-	71.44	75.53	82.83	26.69	30.17	31.92	37.61
	voyage-3-large	-	**81.84**	**85.93**	**90.23**	34.16	40.65	40.92	46.51
	embedding-3-large	-	79.05	82.59	88.58	33.61	38.56	41.03	45.33
	TF-IDF	-	50.13	57.66	64.86	15.81	17.89	18.62	23.71
	BM25	-	48.52	58.41	63.36	15.46	17.84	17.99	22.93
	word2vec	-	55.37	64.89	75.32	12.02	15.05	14.48	21.48
	fastText	-	44.30	50.40	57.52	9.89	12.97	12.40	17.42
	static-similarity-mrl-multilingual-v1	-	31.39	36.55	48.57	8.46	9.81	9.58	13.59
	E5-small-trm-nl	41M	63.57	67.84	76.88	20.03	22.51	24.12	30.17
	mE5-small	118M	60.64	62.98	72.06	21.87	24.43	25.57	30.85
	E5-base-trm-nl	124M	67.98	72.17	78.30	22.32	25.38	28.18	33.13
	potion-multilingual-128M	128M	44.03	53.81	63.27	8.09	9.38	11.04	16.05
	mContriever	178M	53.89	60.96	76.18	9.92	13.94	14.44	20.00
	DPR-XM	277M	32.44	39.54	49.00	6.58	8.77	8.31	12.38
	mE5-base	278M	61.53	69.16	74.86	21.03	24.20	25.24	30.48
NL	mGTE	305M	51.77	62.65	69.72	10.81	12.20	12.72	19.69
	E5-large-trm-nl	355M	68.87	74.67	83.37	22.79	26.02	28.56	33.84
	LaBSE	471M	18.62	28.70	48.36	1.96	2.95	2.44	5.45
	mE5-large	560M	68.01	71.82	76.49	25.64	28.95	31.28	35.94
	mE5-large-instruct	560M	71.37	74.56	81.74	26.02	29.46	32.29	36.97
	BGE-M3	568M	65.75	73.97	82.49	20.88	24.97	28.10	32.00
	snowflake-arctic-embed-l-v2.0	568M	62.76	70.91	78.97	13.10	17.33	17.68	24.63
	jina-embeddings-v3	572M	66.39	73.27	79.96	17.05	20.64	21.17	28.26
	E5-mistral-7b	7B	74.73	78.35	83.67	**36.37**	**40.69**	**42.11**	**46.27**
	BGE-Mult.-Gemma2	9B	79.59	83.13	89.72	31.25	36.29	37.73	43.57
	voyage-2-law	-	74.11	78.43	85.22	25.90	30.43	31.87	37.67
	voyage-3-large	-	76.55	81.95	87.46	32.46	37.80	38.88	43.65
	embedding-3-large	-	**80.94**	**85.32**	**90.52**	27.54	32.64	33.08	40.66

Table 6: Retrieval performance on the French (FR) and Dutch (NL).

Lang	Model	Size	R@10	MAP@10	MAP@100	MRR@10	MRR@100	NDCG@10	NDCG@100
NL	BM25 +	-	25.69	14.55	15.46	16.88	17.84	17.99	22.93
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	36.23	19.50	20.21	23.33	23.96	24.57	27.64
	BGE-reranker-base	0.3B	37.31	19.15	19.80	21.97	22.68	24.33	27.22
	GTE-multilingual-reranker-base	0.3B	32.15	15.73	16.71	19.23	20.14	20.59	24.65
	Jina-reranker-v2-base-multilingual	0.3B	35.67	18.56	19.31	22.06	22.65	23.62	26.82
	BGE-reranker-large	0.6B	38.76	21.81	22.38	25.94	26.42	27.07	29.50
	BGE-reranker-v2-m3	0.6B	**39.11**	**22.34**	**22.92**	**26.96**	**27.43**	**27.54**	**29.93**
	mE5-small +	0.1B	36.98	20.87	21.87	23.41	24.43	25.57	30.85
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	41.93	18.46	19.38	21.44	22.40	24.81	29.23
	BGE-reranker-base	0.3B	34.62	14.05	15.45	16.84	18.23	19.69	25.92
	GTE-multilingual-reranker-base	0.3B	33.44	15.47	16.91	18.00	19.54	20.47	26.93
	Jina-reranker-v2-base-multilingual	0.3B	34.51	16.80	18.13	19.70	21.04	21.90	28.01
	BGE-reranker-large	0.6B	43.84	20.51	21.41	24.30	25.15	27.24	31.20
	BGE-reranker-v2-m3	0.6B	**44.51**	**23.10**	**23.87**	**27.64**	**28.44**	**29.42**	**33.12**
	mE5-base +	0.2B	37.95	69.16	20.06	21.03	24.20	25.24	30.48
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	41.69	20.18	21.22	23.27	24.21	26.14	30.84
	BGE-reranker-base	0.3B	37.39	15.95	17.25	18.85	19.92	21.84	27.53
	GTE-multilingual-reranker-base	0.3B	38.68	16.81	17.98	19.72	20.76	22.76	28.05
	Jina-reranker-v2-base-multilingual	0.3B	37.71	17.33	18.65	19.66	20.87	22.94	28.64
	BGE-reranker-large	0.6B	45.53	21.52	22.42	25.34	26.09	28.36	32.18
	BGE-reranker-v2-m3	0.6B	**47.25**	**24.44**	**25.13**	**29.48**	**30.05**	**31.21**	**34.46**
	mE5-large-instruct +	0.5B	50.83	25.06	26.02	28.66	29.46	32.29	36.97
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	43.73	20.66	21.89	23.99	25.18	27.06	33.33
	BGE-reranker-base	0.3B	34.59	15.06	16.87	17.98	19.76	20.43	28.99
	GTE-multilingual-reranker-base	0.3B	35.34	15.43	17.26	18.23	19.99	20.99	29.36
	Jina-reranker-v2-base-multilingual	0.3B	46.74	19.90	21.12	23.09	24.20	27.23	32.98
	BGE-reranker-large	0.6B	46.29	20.91	22.29	25.55	26.90	28.17	34.21
	BGE-reranker-v2-m3	0.6B	**51.88**	**25.56**	**26.42**	**30.10**	**30.89**	**33.14**	**37.41**
	voyage-3-large +	-	**56.15**	**31.39**	**32.46**	**37.10**	**37.80**	**38.88**	**43.65**
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	48.57	22.00	23.43	25.73	26.88	29.46	36.11
	BGE-reranker-base	0.3B	39.22	16.50	18.47	19.88	21.58	22.77	31.67
	GTE-multilingual-reranker-base	0.3B	38.14	16.58	18.59	20.38	22.07	22.73	31.78
	Jina-reranker-v2-base-multilingual	0.3B	48.14	21.54	23.03	24.88	26.09	28.93	35.70
	BGE-reranker-large	0.6B	43.49	18.74	20.68	23.29	24.83	26.94	33.95
	BGE-reranker-v2-m3	0.6B	51.96	26.80	28.10	32.02	32.99	34.30	40.04
FR	BM25 +	-	37.74	18.36	19.25	21.84	22.65	23.94	28.44
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	39.54	19.05	20.00	21.86	22.85	24.76	29.16
	BGE-reranker-base	0.3B	30.96	12.88	14.26	15.93	17.20	18.19	24.39
	GTE-multilingual-reranker-base	0.3B	41.29	22.49	23.42	25.57	26.44	28.01	31.97
	Jina-reranker-v2-base-multilingual	0.3B	37.84	12.84	13.71	15.78	16.60	19.57	23.99
	BGE-reranker-large	0.6B	36.77	21.10	22.33	25.41	26.67	25.99	31.12
	BGE-reranker-v2-m3	0.6B	**43.95**	**26.18**	**27.01**	**31.46**	**32.09**	**31.92**	**35.24**
	mE5-small +	0.1B	37.41	18.07	19.09	19.42	20.57	23.12	28.10
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	39.65	20.07	20.89	23.60	24.47	25.65	29.92
	BGE-reranker-base	0.3B	26.23	10.41	11.86	11.95	13.60	14.63	21.93
	GTE-multilingual-reranker-base	0.3B	41.58	21.95	22.80	24.97	25.87	27.58	31.57
	Jina-reranker-v2-base-multilingual	0.3B	34.70	12.16	13.33	15.32	16.42	18.37	23.86
	BGE-reranker-large	0.6B	39.86	18.54	19.48	21.66	22.53	24.45	28.74
	BGE-reranker-v2-m3	0.6B	**44.27**	**25.83**	**26.53**	**29.78**	**30.40**	**31.38**	**34.63**
	mE5-base +	0.2B	40.00	19.82	20.71	21.91	22.91	25.37	29.77
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	39.97	20.46	21.33	23.41	24.28	25.94	30.31
	BGE-reranker-base	0.3B	30.12	11.73	12.97	13.37	14.80	16.59	23.09
	GTE-multilingual-reranker-base	0.3B	42.66	22.65	23.43	25.19	25.89	28.31	32.05
	Jina-reranker-v2-base-multilingual	0.3B	36.61	12.54	13.66	15.40	16.45	18.99	24.20
	BGE-reranker-large	0.6B	39.75	20.24	21.19	23.41	24.27	25.73	30.15
	BGE-reranker-v2-m3	0.6B	**45.00**	**26.56**	**27.26**	**30.61**	**31.24**	**32.13**	**35.36**
	mE5-large-instruct +	0.5B	44.56	20.61	21.92	23.83	24.97	27.29	33.87
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	44.16	20.22	21.53	23.11	24.41	26.77	33.59
	BGE-reranker-base	0.3B	27.63	11.15	13.18	12.66	14.74	15.55	25.99
	GTE-multilingual-reranker-base	0.3B	45.67	22.62	23.98	25.06	26.40	29.07	35.66
	Jina-reranker-v2-base-multilingual	0.3B	40.29	13.68	15.33	16.53	18.07	20.81	28.68
	BGE-reranker-large	0.6B	41.69	20.63	22.10	24.37	25.77	26.68	33.99
	BGE-reranker-v2-m3	0.6B	**49.62**	**29.75**	**30.98**	**34.15**	**35.26**	**35.84**	**41.53**
	voyage-3-large +	-	**58.52**	**32.82**	**34.16**	**39.72**	**40.65**	**40.92**	**46.51**
	mmarco-mMiniLMv2-L12-H384-v1	0.1B	46.96	24.14	25.86	27.55	29.11	30.64	38.90
	BGE-reranker-base	0.3B	30.66	10.76	12.92	12.42	14.52	16.09	27.49
	GTE-multilingual-reranker-base	0.3B	47.95	22.77	24.16	24.95	26.21	29.63	37.18
	Jina-reranker-v2-base-multilingual	0.3B	40.59	13.90	15.85	16.91	18.59	21.14	30.62
	BGE-reranker-large	0.6B	41.37	19.49	21.21	23.22	24.64	25.70	34.65
	BGE-reranker-v2-m3	0.6B	51.98	30.15	31.65	35.10	36.23	36.77	43.73

Table 7: Reranker performance on retrieved results from the Dutch (NL) and French (FR) subsets of bLLeQA (test set). First row of each block: retrieval-only baseline; subsequent rows: rerankers applied to the retriever results above.

			French			Dutch		
Model	**Size**	**Setting**	**Pr**	**Rec**	**F1**	**Pr**	**Rec**	**F1**
		Gold	81.94	72.56	75.55	80.00	71.44	74.22
Ministral-3B	3B	RAG+	23.17	27.69	23.52	22.46	29.35	23.51
		RAG	32.22	37.74	33.20	37.60	44.84	38.36
		Gold	83.31	73.40	76.12	69.68	64.60	66.26
Gemma-3-4B-it	4B	RAG+	7.53	39.30	10.85	7.06	37.47	10.14
		RAG	6.89	34.60	9.68	6.75	30.81	9.53
		Gold	98.71	89.85	92.71	98.71	88.99	92.16
Ministral-8B	8B	RAG+	39.03	57.66	42.00	42.91	54.33	44.39
		RAG	37.91	53.95	40.42	38.13	49.35	40.27
		Gold	76.77	68.77	71.21	73.55	65.70	68.24
Qwen3.5-9B	9B	RAG+	43.52	59.22	46.31	42.85	55.66	44.70
		RAG	47.33	58.82	49.37	52.19	61.45	53.58
		Gold	95.27	84.20	87.77	91.83	80.45	84.04
Gemma-3-12B-it	12B	RAG+	18.07	33.63	20.34	15.84	27.31	17.05
		RAG	15.34	29.46	17.04	19.97	33.44	22.36
		Gold	97.58	88.97	91.77	97.63	87.51	90.68
Ministral-14B	14B	RAG+	44.34	56.51	45.28	46.32	53.95	46.01
		RAG	43.61	55.13	45.36	44.46	51.34	44.20
		Gold	80.97	72.81	75.19	74.19	65.58	68.26
GPT-oss-20B	20B	RAG+	33.30	36.24	32.56	33.89	37.50	33.19
		RAG	34.61	36.53	33.77	39.53	42.34	38.45
		Gold	85.16	75.55	78.65	89.35	78.13	81.58
Gemma-3-27B-it	27B	RAG+	28.48	32.37	27.92	26.08	28.17	25.53
		RAG	27.41	30.75	27.14	27.28	27.63	26.39
		Gold	77.42	70.41	72.73	77.42	70.20	72.45
Qwen3.5-27B	27B	RAG+	47.19	72.44	52.54	48.37	73.78	54.00
		RAG	49.37	69.17	53.58	52.02	70.89	56.21
		Gold	96.62	89.63	91.57	93.87	88.12	90.04
GLM-4.7-Flash	30B	RAG+	24.79	58.06	29.81	25.18	62.47	31.06
		RAG	24.20	56.99	29.70	22.77	52.07	27.15
		Gold	96.77	84.93	88.68	98.71	88.11	91.46
Qwen3-30B-A3B-Instruct	30B	RAG+	40.74	54.70	42.29	36.68	57.02	39.33
		RAG	39.46	53.92	41.28	36.97	57.10	40.26
		Gold	71.61	66.54	68.15	74.19	68.31	70.19
Qwen3.5-35B-A3B	35B	RAG+	50.33	64.99	52.91	48.45	62.84	50.71
		RAG	54.98	67.77	57.19	54.69	65.22	56.34
		Gold	90.81	79.37	82.92	90.97	81.27	84.45
Llama-3.3-70B-Instruct	70B	RAG+	36.03	45.38	36.33	35.45	46.72	35.72
		RAG	43.31	50.22	43.24	38.25	44.84	37.68
		Gold	89.68	78.46	81.79	91.61	79.67	83.28
Qwen3-Next-80B-A3B-Instruct	80B	RAG+	46.94	61.15	47.53	44.43	61.56	46.25
		RAG	48.27	58.66	48.24	49.32	60.81	49.88
		Gold	90.32	80.68	83.75	88.17	78.93	81.90
Llama-4-Scout	109B	RAG+	36.73	47.01	37.35	33.82	44.73	34.75
		RAG	39.75	48.25	40.61	36.13	47.02	37.51
		Gold	85.48	76.92	79.65	85.16	76.57	79.24
GPT-oss-120B	120B	RAG+	39.81	50.30	41.25	38.89	47.07	40.06
		RAG	41.49	48.79	42.61	40.09	47.61	41.11
		Gold	70.97	63.48	65.94	65.16	59.35	61.27
Qwen3.5-122B-A10B	122B	RAG+	52.86	71.96	56.60	50.11	64.46	52.35
		RAG	55.94	69.81	58.79	58.20	68.31	60.05
		Gold	95.48	84.53	87.89	93.55	82.95	86.25
Qwen3-235B-A22B	235B	RAG+	48.17	64.60	49.91	48.42	59.65	49.31
		RAG	45.57	59.95	47.33	43.86	53.82	44.30
		Gold	72.26	65.42	67.66	74.19	65.42	68.08
Qwen3.5-397B-A17B	397B	RAG+	50.22	69.59	54.03	50.98	70.05	54.80
		RAG	59.49	76.13	62.39	54.62	66.18	55.88
		Gold	88.82	78.37	81.72	89.68	79.15	82.25
Llama-4-Maverick	400B	RAG+	46.37	55.40	46.60	48.25	63.17	50.01
		RAG	49.64	62.15	51.20	51.34	56.94	50.43
		Gold	95.48	86.95	89.69	93.55	85.12	87.77
Mistral-Large-2512	675B	RAG+	50.19	68.36	53.52	52.38	68.95	55.00
		RAG	46.13	62.37	49.15	47.32	61.77	49.56
		Gold	76.77	69.25	71.60	90.19	81.04	83.97
DeepSeek-v3.2	685B	RAG+	52.89	68.99	55.43	49.52	69.29	52.94
		RAG	53.50	70.24	56.01	48.55	62.63	49.89
		Gold	76.13	68.78	71.13	75.48	68.16	70.49
GLM-5	754B	RAG+	55.52	74.59	58.89	56.49	69.38	58.49
		RAG	59.55	71.21	61.03	58.06	68.41	59.23
		Gold	85.16	72.87	76.78	86.29	75.42	78.61
Kimi-K2-Instruct-0905	1T	RAG+	46.46	56.10	47.20	38.49	46.85	38.83
		RAG	48.95	53.49	48.09	40.00	48.25	40.42
		Gold	64.52	58.48	60.38	66.45	59.53	61.69
Kimi-K2-Thinking	1T	RAG+	46.66	69.72	50.81	43.83	61.51	47.26
		RAG	53.10	72.02	56.82	51.40	68.82	54.70
		Gold	69.68	62.03	64.47	66.45	60.56	62.39
Kimi-K2.5	1.1T	RAG+	54.62	63.59	55.44	54.65	65.19	55.58
		RAG	64.37	71.40	64.49	62.55	68.92	62.55
		Gold	86.45	79.57	81.90	87.10	80.79	82.75
GPT-5-Nano	-	RAG+	33.04	44.09	34.27	28.04	38.03	29.26
		RAG	35.82	44.44	36.18	36.26	45.46	37.06
		Gold	78.71	70.74	73.21	78.06	70.34	72.66
Gemini-3.1-Flash-Lite	-	RAG+	64.62	69.30	63.58	65.64	67.49	63.26
		RAG	63.74	67.04	62.45	63.75	65.65	61.77
		Gold	81.29	79.52	80.18	82.58	79.84	80.80
GPT-5-Mini	-	RAG+	35.74	70.56	43.34	33.93	73.19	42.21
		RAG	32.94	62.12	39.52	37.21	67.66	43.24
		Gold	79.03	70.82	73.49	81.29	74.19	76.43
Gemini-2.5-Flash	-	RAG+	46.94	72.67	51.59	43.28	72.19	49.41
		RAG	46.94	68.84	51.18	41.56	67.74	47.04
		Gold	80.65	76.11	77.49	83.87	79.39	80.94
Gemini-3-Flash	-	RAG+	42.49	76.76	50.31	43.21	78.05	51.74
		RAG	44.25	72.80	50.46	40.16	70.78	46.92
		Gold	77.42	69.33	71.97	76.77	68.68	71.25
Claude-Haiku-4.5	-	RAG+	50.18	71.74	53.78	53.53	67.66	55.87
		RAG	54.05	69.57	56.21	55.62	66.05	56.30
		Gold	78.71	72.03	74.17	79.35	71.66	74.13
Gemini-3.1-Pro	-	RAG+	60.88	69.96	61.51	62.85	71.19	63.19
		RAG	63.88	71.56	63.88	64.28	69.73	63.43
		Gold	70.32	65.74	67.19	70.97	65.59	67.35
GPT-5.4	-	RAG+	41.86	71.63	48.57	41.73	67.33	47.03
		RAG	52.22	75.38	57.49	52.13	74.73	56.73
		Gold	86.45	83.07	84.22	88.39	83.72	85.27
Claude-Sonnet-4.6	-	RAG+	43.60	82.19	52.10	44.88	78.40	52.25
		RAG	42.64	71.26	48.72	41.82	71.61	48.43

Table 8: Citation coverage precision (Pr), recall (Rec), and F1 for Dutch and French, ordered by model size when available.

			French		Dutch	
model	**size**	**setting**	**Correctness**	**Faithfulness**	**Correctness**	**Faithfulness**
		Gold	2.7226	0.7429	2.6258	0.6827
Ministral-3B	3B	RAG+	2.3032	0.6506	2.2774	0.6434
		RAG	2.6323	0.6270	2.7226	0.6094
		Gold	2.6645	0.7820	2.2258	0.6944
Gemma-3-4B-it	4B	RAG+	2.0065	0.4666	1.9355	0.4920
		RAG	2.1226	0.4618	1.8710	0.4938
		Gold	3.3613	0.7506	3.2839	0.7624
Ministral-8B	8B	RAG+	3.4129	0.6930	3.2387	0.7409
		RAG	3.2516	0.7346	3.1871	0.7090
		Gold	2.8323	0.8900	2.6903	0.9257
Qwen3.5-9B	9B	RAG+	3.1226	0.8828	3.0387	0.8661
		RAG	3.3226	0.9004	3.2129	0.8454
		Gold	3.0581	0.9013	2.9548	0.8668
Gemma-3-12B-it	12B	RAG+	2.2194	0.6296	2.0194	0.6462
		RAG	2.2194	0.6292	2.2323	0.6907
		Gold	3.5290	0.7977	3.4000	0.7969
Ministral-14b	14B	RAG+	3.3290	0.7269	3.1548	0.7318
		RAG	3.4387	0.7692	3.1806	0.7486
		Gold	2.8774	0.8646	2.7355	0.7870
GPT-oss-20B	20B	RAG+	2.6581	0.6566	2.3548	0.6552
		RAG	2.8323	0.6675	2.7548	0.6586
		Gold	2.8903	0.9067	2.8903	0.9203
Gemma-3-27B-it	27B	RAG+	2.3871	0.8501	2.1806	0.8190
		RAG	2.3935	0.8384	2.2839	0.8161
		Gold	3.0258	0.9168	2.8968	0.9354
Qwen3.5-27B	27B	RAG+	3.6516	0.9388	3.5097	0.9510
		RAG	3.7032	0.9692	3.6452	0.9344
		Gold	3.3226	0.8679	3.4323	0.8264
Qwen3-30B-A3B-Instruct	30B	RAG+	3.0980	0.7969	2.9605	0.7142
		RAG	3.2013	0.7828	3.0395	0.7673
		Gold	3.2129	0.8561	3.0129	0.8430
GLM-4.7-Flash	30B	RAG+	2.8839	0.7826	2.7613	0.7523
		RAG	2.8516	0.7854	2.7613	0.7608
		Gold	2.8839	0.8832	2.8645	0.9485
Qwen3.5-35B-A3B	35B	RAG+	3.3871	0.9411	3.2323	0.9064
		RAG	3.5742	0.9445	3.5226	0.9013
		Gold	2.6194	0.8874	2.8774	0.8471
Llama-3.3-70B-Instruct	70B	RAG+	2.3613	0.7758	2.2387	0.6843
		RAG	2.4774	0.7978	2.5226	0.7033
		Gold	3.1226	0.9053	2.9742	0.8811
Qwen3-Next-80B-A3B-Instruct	80B	RAG+	3.2774	0.8937	2.9677	0.8304
		RAG	3.2581	0.8652	3.1097	0.8098
		Gold	2.8774	0.8882	2.9032	0.8594
Llama-4-Scout	109B	RAG+	2.4645	0.7488	2.4774	0.6886
		RAG	2.6516	0.7794	2.5548	0.6944
		Gold	3.0452	0.8775	2.9613	0.8825
GPT-oss-120B	120B	RAG+	3.0323	0.7868	2.8000	0.7731
		RAG	3.1548	0.7558	3.1032	0.7764
		Gold	2.8903	0.9599	2.7419	0.9380
Qwen3.5-122B-A10B	122B	RAG+	3.4710	0.9591	3.3548	0.8893
		RAG	3.7161	0.9443	3.6452	0.9419
		Gold	3.4774	0.9083	3.3806	0.8752
Qwen3-235B-A22B	235B	RAG+	3.2323	0.8634	3.1613	0.7978
		RAG	3.1677	0.8647	3.0581	0.8294
		Gold	2.9613	0.9071	2.9419	0.9302
Qwen3.5-397B-A17B	397B	RAG+	3.5806	0.9247	3.5161	0.9485
		RAG	3.8839	0.9671	3.7097	0.9326
		Gold	3.0645	0.9164	2.8774	0.8798
Llama-4-Maverick	400B	RAG+	2.8065	0.8505	2.9548	0.8180
		RAG	3.0974	0.8467	3.0323	0.8160
		Gold	3.5677	0.8749	3.5419	0.8500
Mistral-Large-2512	675B	RAG+	3.7806	0.8549	3.7161	0.8356
		RAG	3.6774	0.8504	3.6774	0.8224
		Gold	2.9806	0.8956	3.1161	0.9068
DeepSeek-v3.2	685B	RAG+	3.4581	0.9288	3.3806	0.8781
		RAG	3.7226	0.9291	3.2903	0.9157
		Gold	3.1484	0.9002	3.0774	0.9430
GLM-5	754B	RAG+	3.6194	0.9593	3.4968	0.9394
		RAG	3.8129	0.9499	3.7161	0.9273
		Gold	3.0714	0.8586	3.0258	0.7778
Kimi-K2-Instruct-0905	1T	RAG+	3.2387	0.7515	2.8323	0.6869
		RAG	3.1935	0.7582	2.9226	0.6774
		Gold	2.9290	0.8520	2.9032	0.9635
Kimi-K2-Thinking	1T	RAG+	3.5548	0.8950	3.3355	0.9009
		RAG	3.8710	0.9278	3.7226	0.8885
		Gold	3.0065	0.9024	2.9032	0.9441
Kimi-K2.5	1.1T	RAG+	3.4258	0.9106	3.3613	0.9313
		RAG	3.8387	0.9380	3.7290	0.9300
		Gold	3.2710	0.9088	3.2774	0.8873
GPT-5-Nano	-	RAG+	3.2000	0.7548	3.0581	0.7276
		RAG	3.2645	0.7894	3.2065	0.7732
		Gold	3.1097	0.9233	3.0516	0.9114
Gemini-3.1-Flash-Lite	-	RAG+	3.4065	0.9466	3.3742	0.8943
		RAG	3.5484	0.9398	3.4581	0.8914
		Gold	3.3548	0.9355	3.3161	0.9466
GPT-5-Mini	-	RAG+	3.5548	0.9547	3.6903	0.9534
		RAG	3.6710	0.9540	3.7032	0.9388
		Gold	3.1806	0.9175	3.2581	0.9223
Gemini-2.5-Flash	-	RAG+	3.7226	0.8942	3.7806	0.9046
		RAG	3.7742	0.9342	3.6710	0.9180
		Gold	3.3355	0.9192	3.3677	0.9536
Gemini-3-Flash	-	RAG+	3.9161	0.9719	3.8387	0.9667
		RAG	3.9935	0.9814	3.8452	0.9573
		Gold	3.1548	0.8949	3.1032	0.9446
Claude-Haiku-4.5	-	RAG+	3.5484	0.9502	3.4065	0.9265
		RAG	3.7742	0.9017	3.6000	0.9010
		Gold	3.1871	0.9391	3.1613	0.9416
Gemini-3.1-Pro	-	RAG+	3.6065	0.9661	3.4645	0.9475
		RAG	3.7355	0.9683	3.7032	0.9220
		Gold	3.1226	0.8907	3.1613	0.9555
GPT-5.4	-	RAG+	3.6194	0.9621	3.5419	0.9515
		RAG	4.0323	0.9475	3.9548	0.9482
		Gold	3.6323	0.9248	3.6129	0.9488
Claude-Sonnet-4.6	-	RAG+	4.1355	0.9714	4.0387	0.9480
		RAG	4.1548	0.9621	4.1290	0.9455

Table 9: Answer correctness and faithfulness results by model, setting, and language.

			French				Dutch			
model	**size**	**setting**	**AccAns**	**InaccAns**	**CorrRef**	**IncorrRef**	**AccAns**	**InaccAns**	**CorrRef**	**IncorrRef**
		Gold	27.74	56.77	0	15.48	29.03	53.55	0	17.42
Ministral-3B	3B	RAG+	21.94	44.52	0	33.55	20.00	50.32	0	29.68
		RAG	17.42	45.16	11.61	25.81	20.00	47.74	12.26	20.00
		Gold	23.87	67.74	0	8.39	14.19	69.68	0	16.13
Gemma-3-4B-it	4B	RAG+	7.74	89.68	0	2.58	7.10	90.97	0	1.94
		RAG	10.97	86.45	1.94	0.65	3.23	91.61	0.65	4.52
		Gold	52.26	47.74	0	0	45.16	54.84	0	0
Ministral-8B	8B	RAG+	52.90	47.10	0	0	49.03	49.03	0	1.94
		RAG	47.10	52.90	0	0	43.23	55.48	0.65	0.65
		Gold	32.90	44.52	0	22.58	27.74	47.10	0	25.16
Qwen3.5-9B	9B	RAG+	41.29	51.61	0	7.10	35.48	55.48	0	9.03
		RAG	37.42	50.32	7.10	5.16	32.26	54.19	7.74	5.81
		Gold	31.61	67.74	0	0.65	30.32	65.16	0	4.52
Gemma-3-12B-it	12B	RAG+	16.77	72.90	0	10.32	10.32	67.74	0	21.94
		RAG	11.61	78.71	1.94	7.74	10.32	67.74	5.81	16.13
		Gold	54.19	44.52	0	1.29	50.32	47.74	0	1.94
Ministral-14b	14B	RAG+	51.61	44.52	0	3.87	47.10	44.52	0	8.39
		RAG	50.32	46.45	2.58	0.65	43.23	50.32	1.94	4.52
		Gold	29.68	52.26	0	18.06	30.32	47.10	0	22.58
GPT-oss-20B	20B	RAG+	27.10	57.42	0	15.48	18.71	59.35	0	21.94
		RAG	23.87	60.00	5.16	10.97	19.35	63.23	6.45	10.97
		Gold	41.94	35.48	0	22.58	36.13	37.42	0	26.45
Qwen3.5-27B	27B	RAG+	61.29	32.26	0	6.45	56.77	35.48	0	7.74
		RAG	56.13	36.13	5.16	2.58	50.97	37.42	5.16	3.87
		Gold	29.68	58.06	0	12.26	27.74	62.58	0	9.68
Gemma-3-27B-it	27B	RAG+	16.77	76.13	0	7.10	11.61	82.58	0	5.81
		RAG	14.84	78.71	3.23	3.23	11.61	83.87	1.94	2.58
		Gold	46.45	50.32	0	3.23	51.61	47.10	0	1.29
Qwen3-30B-A3B-Instruct	30B	RAG+	39.22	57.52	0	3.27	32.24	63.82	0	3.95
		RAG	40.91	55.19	1.95	1.95	37.50	60.53	0.66	1.32
		Gold	36.13	63.87	0	0	36.77	60.00	0	3.23
GLM-4.7-Flash	30B	RAG+	29.03	70.97	0	0	25.16	74.84	0	0
		RAG	31.61	68.39	0	0	21.94	76.13	1.29	0.65
		Gold	37.42	35.48	0	27.10	38.71	36.13	0	25.16
Qwen3.5-35B-A3B	35B	RAG+	54.19	36.13	0	9.68	45.81	41.94	0	12.26
		RAG	49.68	37.42	8.39	4.52	44.52	39.35	9.68	6.45
		Gold	17.42	73.55	0	9.03	26.45	64.52	0	9.03
Llama-3.3-70B-Instruct	70B	RAG+	10.97	78.06	0	10.97	7.10	80.00	0	12.90
		RAG	7.74	76.13	7.74	8.39	13.55	76.77	3.87	5.81
		Gold	37.42	53.55	0	9.03	32.90	58.71	0	8.39
Qwen3-Next-80B-A3B-Instruct	80B	RAG+	46.45	49.03	0	4.52	31.61	63.87	0	4.52
		RAG	38.06	56.77	3.23	1.94	30.97	64.52	3.87	0.65
		Gold	23.87	67.10	0	9.03	32.90	56.13	0	10.97
Llama-4-Scout	109B	RAG+	16.13	65.81	0	18.06	15.48	78.06	0	6.45
		RAG	17.42	69.03	3.87	9.68	13.55	80.65	1.29	4.52
		Gold	34.84	50.97	0	14.19	33.55	51.61	0	14.84
GPT-oss-120B	120B	RAG+	34.19	58.71	0	7.10	27.74	62.58	0	9.68
		RAG	36.77	52.90	4.52	5.81	34.19	54.84	5.81	5.16
		Gold	37.42	34.19	0	28.39	36.77	28.39	0	34.84
Qwen3.5-122B-A10B	122B	RAG+	52.90	38.71	0	8.39	50.97	38.06	0	10.97
		RAG	50.32	40.00	8.39	1.29	50.32	32.90	9.68	7.10
		Gold	54.84	40.65	0	4.52	50.32	43.23	0	6.45
Qwen3-235B-A22B	235B	RAG+	47.10	49.68	0	3.23	40.00	56.77	0	3.23
		RAG	40.00	54.84	2.58	2.58	34.19	60.65	2.58	2.58
		Gold	41.94	30.32	0	27.74	38.06	36.77	0	25.16
Qwen3.5-397B-A17B	397B	RAG+	60.65	29.68	0	9.68	56.77	31.61	0	11.61
		RAG	57.42	31.61	9.03	1.94	52.26	33.55	8.39	5.81
		Gold	32.90	56.77	0	10.32	26.45	63.23	0	10.32
Llama-4-Maverick	400B	RAG+	27.74	58.71	0	13.55	29.03	67.10	0	3.87
		RAG	29.87	57.79	6.49	5.84	27.74	69.03	2.58	0.65
		Gold	57.42	38.71	0	3.87	58.71	34.84	0	6.45
Mistral-Large-2512	675B	RAG+	67.74	32.26	0	0	63.23	36.77	0	0
		RAG	60.65	39.35	0	0	66.45	33.55	0	0
		Gold	43.23	33.55	0	23.23	38.06	52.26	0	9.68
DeepSeek-v3.2	685B	RAG+	56.77	31.61	0	11.61	49.03	46.45	0	4.52
		RAG	50.97	34.84	9.03	5.16	42.58	53.55	2.58	1.29
		Gold	48.39	27.74	0	23.87	44.52	31.61	0	23.87
GLM-5	754B	RAG+	60.00	33.55	0	6.45	56.13	35.48	0	8.39
		RAG	52.90	34.84	9.03	3.23	54.19	36.77	5.81	3.23
		Gold	38.31	53.90	0	7.79	36.13	59.35	0	4.52
Kimi-K2-Instruct-0905	1T	RAG+	43.87	52.90	0	3.23	30.97	58.71	0	10.32
		RAG	40.65	53.55	3.87	1.94	37.42	57.42	1.29	3.87
		Gold	47.74	18.71	0	33.55	44.52	21.94	0	33.55
Kimi-K2-Thinking	1T	RAG+	60.65	27.74	0	11.61	51.61	33.55	0	14.84
		RAG	61.29	25.16	9.68	3.87	52.90	33.55	9.03	4.52
		Gold	44.52	25.81	0	29.68	43.23	23.87	0	32.90
Kimi-K2.5	1.1T	RAG+	56.13	27.74	0	16.13	52.90	29.03	0	18.06
		RAG	58.71	23.23	12.26	5.81	45.81	35.48	12.26	6.45
		Gold	49.68	36.77	0	13.55	46.45	40.65	0	12.90
GPT-5-Nano	-	RAG+	47.74	41.29	0	10.97	41.94	47.10	0	10.97
		RAG	43.87	43.87	3.87	8.39	41.94	45.81	4.52	7.74
		Gold	45.81	34.19	0	20.00	42.58	36.77	0	20.65
Gemini-3.1-Flash-Lite	-	RAG+	49.03	43.23	0	7.74	47.10	48.39	0	4.52
		RAG	47.10	45.81	5.16	1.94	39.35	54.84	0.65	5.16
		Gold	54.84	26.45	0	18.71	53.55	29.03	0	17.42
GPT-5-Mini	-	RAG+	62.58	28.39	0	9.03	66.45	29.03	0	4.52
		RAG	60.65	30.32	3.87	5.16	57.42	34.19	5.81	2.58
		Gold	47.74	31.61	0	20.65	48.39	34.19	0	17.42
Gemini-2.5-Flash	-	RAG+	63.23	32.90	0	3.87	64.52	31.61	0	3.87
		RAG	60.00	35.48	3.87	0.65	59.35	37.42	2.58	0.65
		Gold	55.48	25.81	0	18.71	52.90	30.97	0	16.13
Gemini-3-Flash	-	RAG+	72.90	20.00	0	7.10	70.32	23.23	0	6.45
		RAG	67.10	24.52	5.81	2.58	64.52	29.03	3.87	2.58
		Gold	49.03	28.39	0	22.58	46.45	30.32	0	23.23
Claude-Haiku-4.5	-	RAG+	58.06	34.19	0	7.74	53.55	36.77	0	9.68
		RAG	60.00	30.32	6.45	3.23	50.32	38.06	6.45	5.16
		Gold	46.45	32.90	0	20.65	47.74	32.90	0	19.35
Gemini-3.1-Pro	-	RAG+	56.77	35.48	0	7.74	49.68	41.94	0	8.39
		RAG	49.68	38.71	9.03	2.58	51.61	38.71	8.39	1.29
		Gold	49.68	20.65	0	29.68	51.61	19.35	0	29.03
GPT-5.4	-	RAG+	63.87	20.00	0	16.13	61.94	20.00	0	18.06
		RAG	60.00	18.06	14.19	7.74	58.71	18.06	14.84	8.39
		Gold	62.58	23.87	0	13.55	60.00	28.39	0	11.61
Claude-Sonnet-4.6	-	RAG+	77.42	18.06	0	4.52	75.48	19.35	0	5.16
		RAG	76.77	18.71	2.58	1.94	74.19	18.71	3.87	3.23

Table 10: Share of accurate/inaccurate answers and correct/incorrect refusals (%). AccAns/InaccAns denote accurate/inaccurate answers; CorrRef/IncorrRef denote correct/incorrect refusals. Answers with a correctness score of 4–5 from the LLM judge are considered accurate, while scores of 1–3 are considered inaccurate.

model	size	setting	RefRate	F1-macro	RefPr	RefRec	RefF1	NonRefPr	NonRefRec	NonRefF1
		Gold	17.42	-	-	-	-	100	82.58	90.46
Ministral-3B	3B	RAG+	29.68	-	-	-	-	100	70.32	82.58
		RAG	32.26	56.74	38.00	42.22	40.00	75.24	71.82	73.49
		Gold	16.13	-	-	-	-	100	83.87	91.23
Gemma-3-4B-it	4B	RAG+	1.94	-	-	-	-	100	98.06	99.02
		RAG	5.16	41.96	12.50	2.22	3.77	70.07	93.64	80.16
		Gold	0	-	-	-	-	100	100	100
Ministral-8B	8B	RAG+	1.94	-	-	-	-	100	98.06	99.02
		RAG	1.29	43.57	50.00	2.22	4.26	71.24	99.09	82.89
		Gold	25.16	-	-	-	-	100	74.84	85.61
Qwen3.5-9B	9B	RAG+	9.03	-	-	-	-	100	90.97	95.27
		RAG	13.55	59.58	57.14	26.67	36.36	75.37	91.82	82.79
		Gold	4.52	-	-	-	-	100	95.48	97.69
Gemma-3-12B-it	12B	RAG+	21.94	-	-	-	-	100	78.06	87.68
		RAG	21.94	48.19	26.47	20	22.78	70.25	77.27	73.59
		Gold	1.94	-	-	-	-	100	98.06	99.02
Ministral-14B	14B	RAG+	8.39	-	-	-	-	100	91.61	95.62
		RAG	6.45	45.85	30	6.67	10.91	71.03	93.64	80.78
		Gold	22.58	-	-	-	-	100	77.42	87.27
GPT-oss-20B	20B	RAG+	21.94	-	-	-	-	100	78.06	87.68
		RAG	17.42	52.96	37.04	22.22	27.78	72.66	84.55	78.15
		Gold	9.68	-	-	-	-	100	90.32	94.92
Gemma-3-27B-it	27B	RAG+	5.81	-	-	-	-	100	94.19	97.01
		RAG	4.52	46.85	42.86	6.67	11.54	71.62	96.36	82.17
		Gold	26.45	-	-	-	-	100	73.55	84.76
Qwen3.5-27B	27B	RAG+	7.74	-	-	-	-	100	92.26	95.97
		RAG	11.61	61.15	66.67	26.67	38.1	75.91	94.55	84.21
		Gold	1.29	-	-	-	-	100	98.71	99.35
Qwen3-30B-A3B-Instruct	30B	RAG+	3.87	-	-	-	-	100	96.13	98.03
		RAG	1.94	43.30	33.33	2.22	4.17	71.05	98.18	82.44
		Gold	3.23	-	-	-	-	100	96.77	98.36
GLM-4.7-Flash	30B	RAG+	0	-	-	-	-	100	100	100
		RAG	1.94	45.77	66.67	4.44	8.33	71.71	99.09	83.21
		Gold	25.16	-	-	-	-	100	74.84	85.61
Qwen3.5-35B-A3B	35B	RAG+	12.26	-	-	-	-	100	87.74	93.47
		RAG	16.13	63.10	60	33.33	42.86	76.92	90.91	83.33
		Gold	9.03	-	-	-	-	100	90.97	95.27
Llama-3.3-70B-Instruct	70B	RAG+	12.9	-	-	-	-	100	87.1	93.1
		RAG	9.68	50.40	40	13.33	20	72.14	91.82	80.8
		Gold	8.39	-	-	-	-	100	91.61	95.62
Qwen3-Next-80B-A3B-Instruct	80B	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	4.52	53.79	85.71	13.33	23.08	73.65	99.09	84.5
		Gold	10.97	-	-	-	-	100	89.03	94.2
Llama-4-Scout	109B	RAG+	6.45	-	-	-	-	100	93.55	96.67
		RAG	5.81	43.94	22.22	4.44	7.41	70.55	93.64	80.47
		Gold	14.84	-	-	-	-	100	85.16	91.99
GPT-oss-120B	120B	RAG+	9.68	-	-	-	-	100	90.32	94.92
		RAG	10.97	55.65	52.94	20	29.03	73.91	92.73	82.26
		Gold	34.84	-	-	-	-	100	65.16	78.91
Qwen3.5-122B-A10B	122B	RAG+	10.97	-	-	-	-	100	89.03	94.2
		RAG	16.77	62.55	57.69	33.33	42.25	76.74	90	82.85
		Gold	6.45	-	-	-	-	100	93.55	96.67
Qwen3-235B-A22B	235B	RAG+	3.23	-	-	-	-	100	96.77	98.36
		RAG	5.16	48.79	50	8.89	15.09	72.11	96.36	82.49
		Gold	25.16	-	-	-	-	100	74.84	85.61
Qwen3.5-397B-A17B	397B	RAG+	11.61	-	-	-	-	100	88.39	93.84
		RAG	14.19	60.97	59.09	28.89	38.81	75.94	91.82	83.13
		Gold	10.32	-	-	-	-	100	89.68	94.56
Llama-4-Maverick	400B	RAG+	3.87	-	-	-	-	100	96.13	98.03
		RAG	3.23	49.92	80	8.89	16	72.67	99.09	83.85
		Gold	6.45	-	-	-	-	100	93.55	96.67
Mistral-Large-2512	675B	RAG+	0	-	-	-	-	100	100	100
		RAG	0	41.51	0	0	0	70.97	100	83.02
		Gold	9.68	-	-	-	-	100	90.32	94.92
DeepSeek-v3.2	685B	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	3.87	49.54	66.67	8.89	15.69	72.48	98.18	83.4
		Gold	23.87	-	-	-	-	100	76.13	86.45
GLM-5	754B	RAG+	8.39	-	-	-	-	100	91.61	95.62
		RAG	9.03	57.09	64.29	20	30.51	74.47	95.45	83.67
		Gold	4.52	-	-	-	-	100	95.48	97.69
Kimi-K2-Instruct-0905	1T	RAG+	10.32	-	-	-	-	100	89.68	94.56
		RAG	5.16	44.24	25	4.44	7.55	70.75	94.55	80.93
		Gold	33.55	-	-	-	-	100	66.45	79.84
Kimi-K2-Thinking	1T	RAG+	14.84	-	-	-	-	100	85.16	91.99
		RAG	13.55	63.43	66.67	31.11	42.42	76.87	93.64	84.43
		Gold	32.9	-	-	-	-	100	67.1	80.31
Kimi-K2.5	1.1T	RAG+	18.06	-	-	-	-	100	81.94	90.07
		RAG	18.71	68.05	65.52	42.22	51.35	79.37	90.91	84.75
		Gold	12.9	-	-	-	-	100	87.1	93.1
GPT-5-Nano	-	RAG+	10.97	-	-	-	-	100	89.03	94.2
		RAG	12.26	50.77	36.84	15.56	21.88	72.06	89.09	79.67
		Gold	20.65	-	-	-	-	100	79.35	88.49
Gemini-3.1-Flash-Lite	-	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	5.81	57.39	88.89	17.78	29.63	74.66	99.09	85.16
		Gold	17.42	-	-	-	-	100	82.58	90.46
GPT-5-Mini	-	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	8.39	57.58	69.23	20	31.03	74.65	96.36	84.13
		Gold	17.42	-	-	-	-	100	82.58	90.46
Gemini-2.5-Flash	-	RAG+	3.87	-	-	-	-	100	96.13	98.03
		RAG	3.23	49.92	80	8.89	16	72.67	99.09	83.85
		Gold	16.13	-	-	-	-	100	83.87	91.23
Gemini-3-Flash	-	RAG+	6.45	-	-	-	-	100	93.55	96.67
		RAG	6.45	52.48	60	13.33	21.82	73.1	96.36	83.14
		Gold	23.23	-	-	-	-	100	76.77	86.86
Claude-Haiku-4.5	-	RAG+	9.68	-	-	-	-	100	90.32	94.92
		RAG	11.61	57.17	55.56	22.22	31.75	74.45	92.73	82.59
		Gold	19.35	-	-	-	-	100	80.65	89.29
Gemini-3.1-Pro	-	RAG+	8.39	-	-	-	-	100	91.61	95.62
		RAG	11.61	63.15	72.22	28.89	41.27	76.64	95.45	85.02
		Gold	29.03	-	-	-	-	100	70.97	83.02
GPT-5.4	-	RAG+	18.06	-	-	-	-	100	81.94	90.07
		RAG	23.23	70.75	63.89	51.11	56.79	81.51	88.18	84.72
		Gold	11.61	-	-	-	-	100	88.39	93.84
Claude-Sonnet-4.6	-	RAG+	5.16	-	-	-	-	100	94.84	97.35
		RAG	6.45	52.48	60	13.33	21.82	73.1	96.36	83.14

Table 11: Refusal metrics for Dutch. We report the refusal rate (RefRate) and macro-averaged F1 over refusal versus non-refusal, along with class-wise precision, recall, and F1 for refusals (RefPr/RefRec/RefF1) and non-refusals (NonRefPr/NonRefRec/NonRefF1). "-" indicates undefined metrics.

model	size	setting	RefRate	F1-macro	RefPr	RefRec	RefF1	NonRefPr	NonRefRec	NonRefF1
		Gold	15.48	-	-	-	-	100	84.52	91.61
Ministral-3B	3B	RAG+	33.55	-	-	-	-	100	66.45	79.84
		RAG	37.42	51.29	31.03	40	34.95	72.16	63.64	67.63
		Gold	8.39	-	-	-	-	100	91.61	95.62
Gemma-3-4B-it	4B	RAG+	2.58	-	-	-	-	100	97.42	98.69
		RAG	2.58	47.88	75	6.67	12.24	72.19	99.09	83.52
		Gold	0	-	-	-	-	100	100	100
Ministral-8B	8B	RAG+	0	-	-	-	-	100	100	100
		RAG	0	41.51	0	0	0	70.97	100	83.02
		Gold	0.65	-	-	-	-	100	99.35	99.68
Gemma-3-12B-it	12B	RAG+	10.32	-	-	-	-	100	89.68	94.56
		RAG	9.68	44.2	20	6.67	10	70	89.09	78.4
		Gold	22.58	-	-	-	-	100	77.42	87.27
Qwen3.5-9B	9B	RAG+	7.1	-	-	-	-	100	92.9	96.32
		RAG	12.26	58.65	57.89	24.44	34.38	75	92.73	82.93
		Gold	1.29	-	-	-	-	100	98.71	99.35
Ministral-14B	14B	RAG+	3.87	-	-	-	-	100	96.13	98.03
		RAG	3.23	49.92	80	8.89	16	72.67	99.09	83.85
		Gold	18.06	-	-	-	-	100	81.94	90.07
GPT-oss-20B	20B	RAG+	15.48	-	-	-	-	100	84.52	91.61
		RAG	16.13	50.18	32	17.78	22.86	71.54	84.55	77.5
		Gold	12.26	-	-	-	-	100	87.74	93.47
Gemma-3-27B-it	27B	RAG+	7.1	-	-	-	-	100	92.9	96.32
		RAG	6.45	50.27	50	11.11	18.18	72.41	95.45	82.35
		Gold	22.58	-	-	-	-	100	77.42	87.27
Qwen3.5-27B	27B	RAG+	6.45	-	-	-	-	100	93.55	96.67
		RAG	7.74	55.93	66.67	17.78	28.07	74.13	96.36	83.79
		Gold	3.23	-	-	-	-	100	96.77	98.36
Qwen3-30B-A3B-Instruct	30B	RAG+	3.23	-	-	-	-	100	96.77	98.36
		RAG	3.87	47.2	50	6.67	11.76	71.81	97.27	82.63
		Gold	0	-	-	-	-	100	100	100
GLM-4.7-Flash	30B	RAG+	0	-	-	-	-	100	100	100
		RAG	0	41.51	0	0	0	70.97	100	83.02
		Gold	27.1	-	-	-	-	100	72.9	84.33
Qwen3.5-35B-A3B	35B	RAG+	9.68	-	-	-	-	100	90.32	94.92
		RAG	12.9	62.04	65	28.89	40	76.3	93.64	84.08
		Gold	9.03	-	-	-	-	100	90.97	95.27
Llama-3.3-70B-Instruct	70B	RAG+	10.97	-	-	-	-	100	89.03	94.2
		RAG	16.13	57.56	48	26.67	34.29	74.62	88.18	80.83
		Gold	9.03	-	-	-	-	100	90.97	95.27
Qwen3-Next-80B-A3B-Instruct	80B	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	5.16	51.07	62.5	11.11	18.87	72.79	97.27	83.27
		Gold	9.03	-	-	-	-	100	90.97	95.27
Llama-4-Scout	109B	RAG+	18.06	-	-	-	-	100	81.94	90.07
		RAG	13.55	48.03	28.57	13.33	18.18	70.9	86.36	77.87
		Gold	14.19	-	-	-	-	100	85.81	92.36
GPT-oss-120B	120B	RAG+	7.1	-	-	-	-	100	92.9	96.32
		RAG	10.32	52.04	43.75	15.56	22.95	72.66	91.82	81.12
		Gold	28.39	-	-	-	-	100	71.61	83.46
Qwen3.5-122B-A10B	122B	RAG+	8.39	-	-	-	-	100	91.61	95.62
		RAG	9.68	64.87	86.67	28.89	43.33	77.14	98.18	86.4
		Gold	4.52	-	-	-	-	100	95.48	97.69
Qwen3-235B-A22B	235B	RAG+	3.23	-	-	-	-	100	96.77	98.36
		RAG	5.16	48.79	50	8.89	15.09	72.11	96.36	82.49
		Gold	27.74	-	-	-	-	100	72.26	83.9
Qwen3.5-397B-A17B	397B	RAG+	9.68	-	-	-	-	100	90.32	94.92
		RAG	10.97	65.73	82.35	31.11	45.16	77.54	97.27	86.29
		Gold	10.32	-	-	-	-	100	89.68	94.56
Llama-4-Maverick	400B	RAG+	13.55	-	-	-	-	100	86.45	92.73
		RAG	12.26	56.68	52.63	22.22	31.25	74.26	91.82	82.11
		Gold	3.87	-	-	-	-	100	96.13	98.03
Mistral-Large-2512	675B	RAG+	0	-	-	-	-	100	100	100
		RAG	0	41.51	0	0	0	70.97	100	83.02
		Gold	23.23	-	-	-	-	100	76.77	86.86
DeepSeek-v3.2	685B	RAG+	11.61	-	-	-	-	100	88.39	93.84
		RAG	14.19	62.87	63.64	31.11	41.79	76.69	92.73	83.95
		Gold	23.87	-	-	-	-	100	76.13	86.45
GLM-5	754B	RAG+	6.45	-	-	-	-	100	93.55	96.67
		RAG	12.26	64.56	73.68	31.11	43.75	77.21	95.45	85.37
		Gold	7.74	-	-	-	-	100	92.26	95.97
Kimi-K2-Instruct-0905	1T	RAG+	3.23	-	-	-	-	100	96.77	98.36
		RAG	5.81	52.91	66.67	13.33	22.22	73.29	97.27	83.59
		Gold	33.55	-	-	-	-	100	66.45	79.84
Kimi-K2-Thinking	1T	RAG+	11.61	-	-	-	-	100	88.39	93.84
		RAG	13.55	65.35	71.43	33.33	45.45	77.61	94.55	85.25
		Gold	29.68	-	-	-	-	100	70.32	82.58
Kimi-K2.5	1.1T	RAG+	16.13	-	-	-	-	100	83.87	91.23
		RAG	18.06	68.64	67.86	42.22	52.05	79.53	91.82	85.23
		Gold	13.55	-	-	-	-	100	86.45	92.73
GPT-5-Nano	-	RAG+	10.97	-	-	-	-	100	89.03	94.2
		RAG	12.26	48.81	31.58	13.33	18.75	71.32	88.18	78.86
		Gold	20	-	-	-	-	100	80	88.89
Gemini-3.1-Flash-Lite	-	RAG+	7.74	-	-	-	-	100	92.26	95.97
		RAG	7.1	56.41	72.73	17.78	28.57	74.31	97.27	84.25
		Gold	18.71	-	-	-	-	100	81.29	89.68
GPT-5-Mini	-	RAG+	9.03	-	-	-	-	100	90.97	95.27
		RAG	9.03	50.81	42.86	13.33	20.34	72.34	92.73	81.27
		Gold	20.65	-	-	-	-	100	79.35	88.49
Gemini-2.5-Flash	-	RAG+	3.87	-	-	-	-	100	96.13	98.03
		RAG	4.52	53.79	85.71	13.33	23.08	73.65	99.09	84.5
		Gold	18.71	-	-	-	-	100	81.29	89.68
Gemini-3-Flash	-	RAG+	7.1	-	-	-	-	100	92.9	96.32
		RAG	8.39	57.58	69.23	20	31.03	74.65	96.36	84.13
		Gold	22.58	-	-	-	-	100	77.42	87.27
Claude-Haiku-4.5	-	RAG+	7.74	-	-	-	-	100	92.26	95.97
		RAG	9.68	58.67	66.67	22.22	33.33	75	95.45	84
		Gold	20.65	-	-	-	-	100	79.35	88.49
Gemini-3.1-Pro	-	RAG+	7.74	-	-	-	-	100	92.26	95.97
		RAG	11.61	65.14	77.78	31.11	44.44	77.37	96.36	85.83
		Gold	29.68	-	-	-	-	100	70.32	82.58
GPT-5.4	-	RAG+	16.13	-	-	-	-	100	83.87	91.23
		RAG	21.94	70.27	64.71	48.89	55.7	80.99	89.09	84.85
		Gold	13.55	-	-	-	-	100	86.45	92.73
Claude-Sonnet-4.6	-	RAG+	4.52	-	-	-	-	100	95.48	97.69
		RAG	4.52	49.17	57.14	8.89	15.38	72.3	97.27	82.95

Table 12: Refusal metrics for French. We report the refusal rate (RefRate) and macro-averaged F1 over refusal versus non-refusal, along with class-wise precision, recall, and F1 for refusals (RefPr/RefRec/RefF1) and non-refusals (NonRefPr/NonRefRec/NonRefF1). "-" indicates undefined metrics.

Model	Size	Source
TF-IDF	-	Sparck Jones (1972)
BM25	-	Robertson et al. (1994)
word2vec	-	Mikolov et al. (2013b,a); Tulkens et al. (2016); Fauconnier (2015)
fastText	-	Bojanowski et al. (2017); Grave et al. (2018)
static-similarity-mrl-multilingual-v1	-	Reimers and Gurevych (2019)
E5-small-trm-nl	0.04B	Lotfi et al. (2025a)
mE5-small	0.1B	Wang et al. (2024b)
E5-base-trm-nl	0.1B	Lotfi et al. (2025a)
potion-multilingual-128M	0.1B	Tulkens and van Dongen (2024)
mContriever	0.2B	Izacard et al. (2021)
DPR-XM	0.3B	Louis et al. (2025)
mE5-base	0.3B	Wang et al. (2024b)
mGTE	0.3B	Zhang et al. (2024a)
E5-large-trm-nl	0.4B	Lotfi et al. (2025a)
LaBSE	0.5B	Feng et al. (2022)
mE5-large	0.6B	Wang et al. (2024b)
mE5-large-instruct	0.6B	Wang et al. (2024b)
BGE-M3	0.6B	Chen et al. (2024a)
snowflake-arctic-embed-l-v2.0	0.6B	Yu et al. (2024)
jina-embeddings-v3	0.6B	Sturua et al. (2025)
E5-mistral-7b	7B	Wang et al. (2024a, 2022)
BGE-Mult.-Gemma2	9B	Chen et al. (2024a); Xiao et al. (2024)
voyage-2-law	-	VoyageAI (2024)
voyage-3-large	-	VoyageAI (2025)
embedding-3-large	-	OpenAI (2025b)
mmarco-mMiniLMv2-L12-H384-v1	0.1B	Wang et al. (2021)
BGE-reranker-base	0.3B	Xiao et al. (2024)
GTE-multilingual-reranker-base	0.3B	Zhang et al. (2024b)
Jina-reranker-v2-base-multilingual	0.3B	JinaAI (2025)
BGE-reranker-large	0.6B	Xiao et al. (2024)
BGE-reranker-v2-m3	0.6B	Li et al. (2023); Chen et al. (2024b)
Ministral-3B	3B	Liu et al. (2026)
Gemma-3-4B-it	4B	Team et al. (2025a)
Ministral-8B	8B	Liu et al. (2026)
Qwen3.5-9B	9B	QwenTeam (2026)
Gemma-3-12B-it	12B	Team et al. (2025a)
Ministral-14B	14B	Liu et al. (2026)
GPT-oss-20B	20B	OpenAI (2025a)
Gemma-3-27B-it	27B	Team et al. (2025a)
GLM-4.7-Flash	30B	Team et al. (2025b)
Qwen3-30B-A3B-Instruct-2507	30B	Team (2025)
Qwen3.5-35B-A3B	35B	QwenTeam (2026)
Llama-3.3-70B-Instruct	70B	Meta (2025)
Qwen3-Next-80B-A3B-Instruct	80B	Team (2025); Yang et al. (2025)
Llama-4-Scout	109B	Meta (2026)
GPT-oss-120B	120B	OpenAI (2025a)
Qwen3.5-122B-A10B	122B	QwenTeam (2026)
Qwen3-235B-A22B-Instruct-2507	235B	Team (2025); Yang et al. (2025)
Qwen3.5-122B-A10B	122B	QwenTeam (2026)
Qwen3.5-397B-A17B	397B	QwenTeam (2026)
Llama-4-Maverick	400B	Meta (2026)
Mistral-Large-2512	675B	Mistral AI (2025)
DeepSeek-v3.2	685B	DeepSeek-AI (2025)
GLM-5	754B	GLM-5-Team et al. (2026)
Kimi-K2-Instruct-0905	1T	Team et al. (2026)
Kimi-K2-Thinking	1T	Team et al. (2026)
Kimi-K2.5	1.1T	Moonshot AI (2026)
GPT-5-Nano	-	OpenAI (2025)
Gemini-3.1-Flash-Lite	-	DeepMind (2026)
GPT-5-Mini	-	OpenAI (2025)
Gemini-2.5-Flash	-	DeepMind (2025a)
Gemini-3-Flash	-	DeepMind (2025b)
Claude-Haiku-4.5	-	Anthropic (2025)
Gemini-3.1-Pro	-	DeepMind (2025c)
GPT-5.4	-	OpenAI (2025c)
Claude-Sonnet-4.6	-	Anthropic (2026)

Table 13: Models used for RAG, ordered by model size when available. The sections correspond to retrieval, reranking, and generative models, respectively.

Prompt Text
You are a Senior Legal Evaluator and expert attorney. Your job is to evaluate the quality of a response generated by an AI Legal Assistant by comparing it against a verified "Gold Standard" answer. You will be provided with: 1. A Legal Question. 2. A Gold Standard Answer (created by human experts). 3. A System Response (generated by the AI). You must rate the System Response on a scale of 1 to 5 based on how well the legal reasoning and factual content align with the Gold Standard. - **Content Focus:** Focus ONLY on the textual content, legal accuracy, and reasoning. - **Strict Adherence:** The Gold Standard is the absolute truth. If the System Response contradicts the Gold Standard, it is wrong, even if you believe the Gold Standard might be incomplete. - **completeness:** The System Response must contain the critical legal elements present in the Gold Standard. Assess the response below and return a JSON object with two fields: - `"score"`: an integer between 1 and 5. 1. ** Critical Failure / Incorrect** The response implies the opposite legal conclusion to the Gold Standard, provides dangerous legal advice, or is completely irrelevant to the question. 2. **Poor / Significant Omissions** The conclusion is vague or partially incorrect. It misses the central legal argument or key fact found in the Gold Standard. It may contain hallucinations. 3. **Acceptable / Partially Complete** The response captures the general legal principle correctly but misses important nuances, exceptions, or specific details present in the Gold Standard. It is legally safe but not comprehensive. 4. **Good / Mostly Accurate** The response aligns with the Gold Standard in conclusion and reasoning. It may miss very minor details that do not alter the legal outcome. 5. **Excellent / Semantically Equivalent** The response is logically and factually equivalent to the Gold Standard. It captures all key legal elements, reasoning, and conclusions. (Difference in wording or structure is acceptable). - `"reason"`: a brief explanation for why the score was given. This must mention specific strengths or shortcomings, referencing relevant details from the input. Do **not** quote the score itself in the explanation. Your explanation should: - {reasoning_expectation} - Mention key details from the test case parameters. - Be concise, clear, and focused on the evaluation logic. Only return valid JSON. Do **not** include any extra commentary or text. — Test Case: {test_case_content} Parameters: {parameters} — **Example JSON:** {{ "reason": "your concise and informative reason here", "score": 1 }} JSON: """

Table 14: DeepEval G-Eval prompt for answer correctness.

Language	**Prompt**
French	Étant donné une question juridique, récupére les documents qui peuvent aider à y répondre
Dutch	Gegeven een juridische vraag, haal documenten op die kunnen helpen bij het beantwoorden van de vraag

Table 15: Prompts used for `E5-large-instruct`, `BGE-Gemma2` and `E5-mistral` for the retrieval task.

System Prompt
You are an expert legal assistant specializing in Belgian law. Your task is to answer legal questions to the best of your knowledge of Belgian law. You respond exclusively in $answer_language and prioritize legal accuracy.
User Prompt
Instructions: 1. The legal context is provided as a JSON array of articles. Each article has the following structure: - "id": the unique identifier of the article - "text": the text of the article excerpt 2. Carefully analyze all articles in the legal context and assess their relevance to the legal question. 3. Answer the question ONLY IF the context is sufficient: - The context must contain all necessary rules or conditions to answer the question. - If any essential condition is missing, unclear, or cannot be derived from the provided texts, do not answer. - If relevant articles conflict on a key condition and the conflict cannot be resolved using only the context, do not answer. - Use ONLY the "text" fields. Do not rely on external knowledge or assumptions. 4. Output format requirements: - Return a JSON array of objects: ["text": "...", "supported_sources": ["id1","id2"], ...] - Each object represents exactly one answer paragraph. - EVERY paragraph must be directly supported by one or more article IDs. - "supported_sources" must be a valid JSON array of strings (double quotes). - Include ONLY article IDs that appear in the provided legal context. - Include ONLY IDs that directly support the corresponding paragraph text. - Do not include irrelevant or speculative citations. 5. If the context is insufficient, incomplete, contradictory, or irrelevant, return exactly: ["text": "Insufficient context", "supported_sources": []] Legal question: $question Regions involved: $regions Topics: $topics Legal context (article excerpts): $context Output the JSON array immediately. Do not include any preamble.

Table 16: System and user prompts used by the LLMs in the RAG experiments.