Bilingual BSARD: Extending Statutory Article Retrieval to Dutch

Ehsan Lotfi[*] **Nikolay Banar**[*] **Nerses Yuzbashyan** **Walter Daelemans**

CLiPS, University of Antwerp, Belgium
{ehsan.lotfi, nicolae.banari, nerses.yuzbashyan, walter.daelemans}
@uantwerpen.be

Abstract

Statutory article retrieval plays a crucial role in making legal information more accessible to both laypeople and legal professionals. Multilingual countries like Belgium present unique challenges for retrieval models due to the need for handling legal issues in multiple languages. Building on the Belgian Statutory Article Retrieval Dataset (BSARD, Louis and Spanakis (2022)) in French, we introduce the bilingual version of this dataset, bBSARD. The dataset contains parallel Belgian statutory articles in both French and Dutch, along with legal questions from BSARD and their Dutch translation. Using bBSARD, we conduct extensive benchmarking of retrieval models available for Dutch and French. Our benchmarking setup includes lexical models, zero-shot dense models, and fine-tuned small foundation models. Our experiments show that BM25 remains a competitive baseline compared to many zero-shot dense models in both languages. We also observe that while proprietary models outperform open alternatives in the zero-shot setting, they can be matched or surpassed by fine-tuning small language-specific models. Our dataset and evaluation code are publicly available.

1 Introduction

Open access to legal information is considered a fundamental right according to the Charter of Fundamental Rights in the European Union (European Union, 2012). Effective retrieval models are an essential component to ensuring this right, as they allow laypeople and legal professionals to efficiently search through vast amounts of legal information. In countries like Belgium, where laws are available in multiple languages (e.g. French and Dutch), the need for high-performance legal retrieval models becomes even more crucial, as they require equal accessibility to relevant legal material regardless of the language in use.

The retrieval task (Thakur et al., 2021) has experienced a significant boost due to the recent advances in textual embeddings, which rely on extensively pre-trained large language models (LLMs; Zhao et al., 2024). These models can encode text into vector representations which perform very well across a broad range of tasks (Muennighoff et al., 2023), including classification (Maas et al., 2011; Saravia et al., 2018; O'Neill et al., 2021) and clustering (Aggarwal and Zhai, 2012; Geigle et al., 2021). Open-source models like E5 (Wang et al., 2022, 2023, 2024) and BGE-M3 (Chen et al., 2024), along with private models from VoyageAI (2024) and OpenAI (2024) have shown remarkable results in zero-shot retrieval, across multiple languages, and different domains, including various legal benchmarks (Muennighoff et al., 2023). These developments offer great opportunities to improve accessibility in multilingual legal jurisdictions.

Belgium invests significant resources to consolidate[1] its laws in both French and Dutch, which is done by the manual labor of qualified legal professionals. This results in a highly valuable resource for research in multilingual legal retrieval models. Building on this resource, and the Belgian Statutory Article Retrieval Dataset (BSARD; Louis and Spanakis, 2022) in French, we introduce the Bilingual Belgian Statutory Article Retrieval Dataset (bBSARD), which we curated by scraping parallel Dutch and French articles, and translating the BSARD questions into Dutch. Using bBSARD, we conducted extensive benchmarking of retrieval models available for Dutch and French, both in zero-shot and fine-tuned scenarios.

In addition to a parallel bilingual legal corpus, bBSARD offers a much-needed retrieval benchmark for the Dutch language, allowing for more accurate and reliable evaluation of Dutch retrieval

[*] Indicates equal contribution

[1] https://www.ejustice.just.fgov.be/cgi_loi/contenu.pl?language=nl&view_numac=2019050815nl

Proceedings of the 31st International Conference on Computational Linguistics, pages 10–21
January 19–24, 2025. ©COLING 2025

models. bBSARD dataset and evaluation code are available on the HuggingFace hub[2] (under the `cc-by-nc-sa-4.0` license), and our GitHub repository[3] (under the MIT license), respectively.

2 Related Work

In the last few years, the field of legal NLP has gained increased interest, leading to the development of a growing number of datasets for research. In this section, we focus specifically on datasets that address the task of legal retrieval grounded in legal provisions, including documents, statutory articles, and cases.

CAIL2018 (Xiao et al., 2018; Zhong et al., 2018) is a dataset designed for legal judgment prediction in Chinese, released as part of the Chinese AI and Law Challenge[4]. It contains over 2.68 million Chinese criminal cases, linked to 183 law articles and 202 charges. One of the subtasks from this challenge involved predicting relevant law articles based on the factual descriptions of specific cases. Following this, the CAIL2019-SCM dataset (Xiao et al., 2019) focuses on similar case matching with 8,964 case triplets (in which two cases are similar) sourced from the Supreme People's Court of China.

Zhong et al. (2020) released JEC-QA, a question answering dataset based on the Chinese bar exam. The dataset contains 26,365 multiple-choice questions, along with 3,382 Chinese legal provisions.

The AILA competitions (Bhattacharya et al., 2019, 2021) introduced datasets for precedent and statute retrieval from Indian law, with content in English. For each year, around 50 queries were linked to relevant documents in retrieval corpora containing 197 statutes and around 3,000 prior cases.

Similarly, COLIEE (Rabelo et al., 2021, 2022; Kim et al., 2022; Goebel et al., 2024) competitions include the task of statute article retrieval from provided datasets. For each year, the datasets contain around 100 test questions from the Japanese legal bar exams, labeled with relevant articles from the Japanese Civil Code, translated into English. The provided training sets include up to 1000 question-article pairs.

Chen et al. (2023) introduced EQUALS, a dataset containing 6,914 question-article-answer triplets, with a corresponding retrieval corpus of 3,081 Chinese law articles. The question-answer pairs were collected from a free Chinese legal advice forum, then revised and further annotated by senior law students. Similarly, STARD (Su et al., 2024) introduced 1,543 queries from the general public, with a retrieval corpus of 55,348 Chinese statutory articles.

GerLayQA (Büttner and Habernal, 2024) consists of around 21,000 legal questions from laymen paired with answers from legal professionals and grounded in paragraphs from German law books.

Most related to our work is BSARD (Louis and Spanakis, 2022); a statutory article retrieval dataset which contains over 1,100 legal questions from Belgian citizens and around 22,600 Belgian law articles as the retrieval corpus. LLeQA (Louis et al., 2024b) complements BSARD with answers from legal experts, along with an additional 760 legal questions and 5,308 statutory articles. While LLeQA is a more extensive resource than BSARD, the latter is available under less restrictive terms[5] and does not require a separate user agreement[6].

The resources presented above support the training, evaluation, and benchmarking of retrieval models across different legal domains and languages, highlighting the need for tailored approaches in each jurisdiction. Contributing to the growing field of legal NLP, we introduce bBSARD, a bilingual dataset built on BSARD which offers parallel Belgian law articles in both French and Dutch, along with legal questions translated from French to Dutch. In addition to providing a reliable benchmark for the retrieval task in Dutch, bBSARD aims to address challenges of legal retrieval in multilingual jurisdictions.

3 Dataset

As mentioned, we base our work on the BSARD dataset (Louis and Spanakis, 2022), extending it to the Dutch language by adding the corresponding articles and questions. We discuss the procedure in the following sections.

3.1 Legislation in Dutch

To get the BSARD legislation and articles in Dutch, we leverage **Justel**[7], the multilingual database

[2] https://huggingface.co/datasets/clips/bBSARD

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

[4] http://cail.cipsc.org.cn

[5] https://huggingface.co/datasets/maastrichtlawtech/bsard

[6] https://huggingface.co/datasets/maastrichtlawtech/lleqa

[7] https://www.ejustice.just.fgov.be/cgi_loi/welcome.pl?language=nl

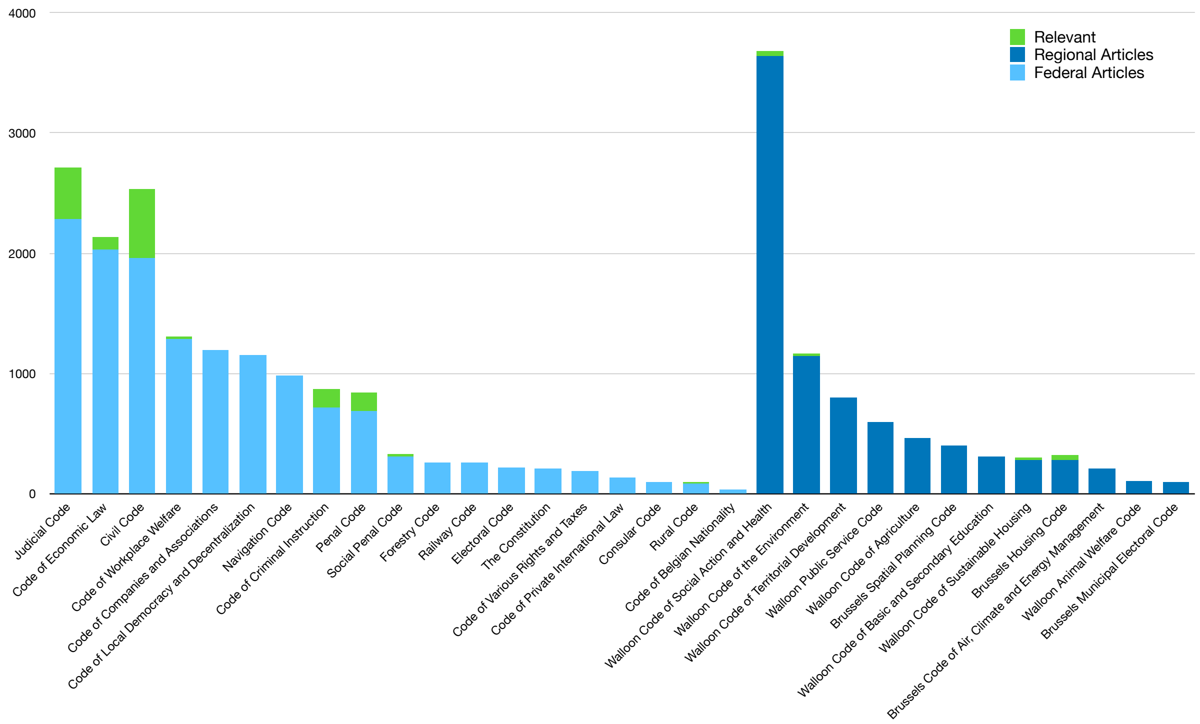

Figure 1: Distribution of different codes in the bBSARD article corpus. ’Relevant’ articles (green) are the ones cited in the question set. Light and dark blue columns correspond to the Federal and Regional codes, respectively.

maintained by the Belgian Federal Government that provides online access to most Belgian legislation in French, Dutch and (often) German. Since there are no public APIs, we scrape the appropriate French and Dutch pages (52 pages for each language), according to the BSARD corpus. Considering the continuous changes and updates, and the fact that BSARD was curated in May 2021 (Louis and Spanakis, 2022), we make sure that the Dutch and French articles come from the same legislative version, by manually controlling their enforcement dates.

In the end, we manage to retrieve and align 22,417 out of 22,633 articles (99%) in both languages (see Appendix A for the alignment process.). The missing 216 articles mostly belong to the Walloon Code of Environment-Decrees (126 articles absent from the Dutch page), and the Military Penal Code (66 articles absent from the database). Fortunately, these missing articles contribute only marginally to the relevant subset (only 1 missing article is cited in a multi-referenced question). Table 3 in Appendix A summarizes the differences between the original and bilingual datasets.

Figure 1 shows how different codes contribute to the complete and relevant set of articles. The majority of relevant articles (i.e. annotated as necessary to answer questions, colored light green in the chart) come from four Federal codes: Judicial, Civil, Penal, and Criminal Instruction.

3.2 Questions in Dutch

BSARD contains 1,108 questions (split as 886/222 for the train/test sets), each labeled by experts with the IDs of the corresponding relevant law articles from the corpus. These questions have been curated in partnership with Droits Quotidiens[8], from emails sent by Belgian citizens to this organization, asking for advice on legal issues. They cover a wide range of topics, with around 85% of them being either about family, housing, money, or justice, while the remaining 15% concern either social security, foreigners, or work (Louis and Spanakis, 2022).

To produce these questions in Dutch, we opt for automatic translation followed by human inspection. We first prompt OpenAI’s GPT-4o with the original French question, as well as the relevant articles (to provide context), and ask for the Dutch translation (The full prompt can be found in Appendix B). To increase translation fidelity, we set the temperature to 0 (Peng et al., 2023). We then asked a native speaker to examine a random sample of 100 translated questions, and annotate them for potential issues. The results showed (legally)

[8] https://droitsquotidiens.be/

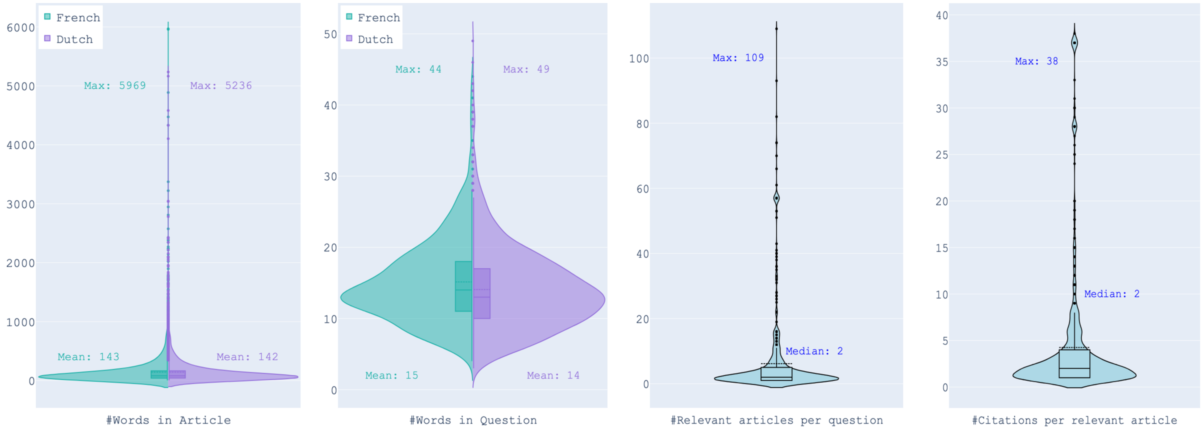

Figure 2: Basic statistics of bBSARD. From the left: Number of words in the articles (French and Dutch), Number of words in the questions (French and Dutch), number of relevant articles per question, and number of citations per relevant article.

inaccurate choice of words in 2%, and minor semantic/grammatical/lexical issues (e.g. translation being too literal) in 6% of the studied samples.

Figure 2 shows basic statistical features of the bBSARD dataset. The French and Dutch articles have an average length of 143 and 142 words, respectively, while for the questions these numbers stand at 15 and 14 words. Regarding the question-article mapping, 1,611 distinct articles (out of 22,417) contribute to the relevant subset, and 75% of questions have fewer than five references, with a median value of two.

4 Experimental Setup

This section describes our experimental setup used to benchmark the retrieval performance of a selection of models on bBSARD. We mostly reuse the codebase from BSARD (Louis and Spanakis, 2022), making modifications where necessary to accommodate the retrieval models to the specific requirements of our experiments. Below we describe the models, data processing steps, and evaluation metrics used in our experiments.

4.1 Models

We select a diverse range of models in three different categories/settings: lexical, zero-shot, and fine-tuned.

4.1.1 Lexical models

Lexical approaches for retrieval rely on keyword matching and utilize various word (or token) weighting schemes and algorithms to determine the relevance of documents for a given query. The most popular algorithms are TF-IDF (Term Frequency-Inverse Document Frequency; Sparck Jones, 1972; Salton and Yang, 1973) and BM25 (Best Match 25; Robertson et al., 1994). Despite the lexical gap issues, where the vocabulary used in queries can differ from that in relevant documents, BM25 remains a robust baseline for many retrieval tasks. Remarkably BM25 was outperformed only recently by E5 (Wang et al., 2022) on the BEIR retrieval benchmark (Thakur et al., 2021) in a setup that does not utilize any labeled data. In our experiments, we evaluate both TF-IDF and BM25.

4.1.2 Zero-shot models

Recently, LLMs achieved impressive results on various retrieval tasks (Zhao et al., 2024). For the zero-shot setting, we select the following multilingual retrieval models, from both open and proprietary categories: mContriever[9] (Izacard et al., 2022), LaBSE (Feng et al., 2022), mE5 (Wang et al., 2024), E5$_{mistral-7b}$ (Wang et al., 2023), BGE-M3 (Chen et al., 2024), DPR-XM (Louis et al., 2024a), BGE-Multilingual-Gemma2 (Li et al., 2024), jina-embeddings-v3 (Sturua et al., 2024), mGTE (Zhang et al., 2024), voyage-3 (VoyageAI, 2024), text-embedding-3-large (OpenAI, 2024). For models with a maximum input length of 512 tokens (except LaBSE), we divide the text into overlapping chunks of 200 tokens with an overlap of 20 tokens between neighboring chunks to mitigate the input length limitations. We do not impose any limits on the input length for other models, allowing them to handle truncation if necessary.

[9] https://huggingface.co/facebook/mcontriever-msmarco

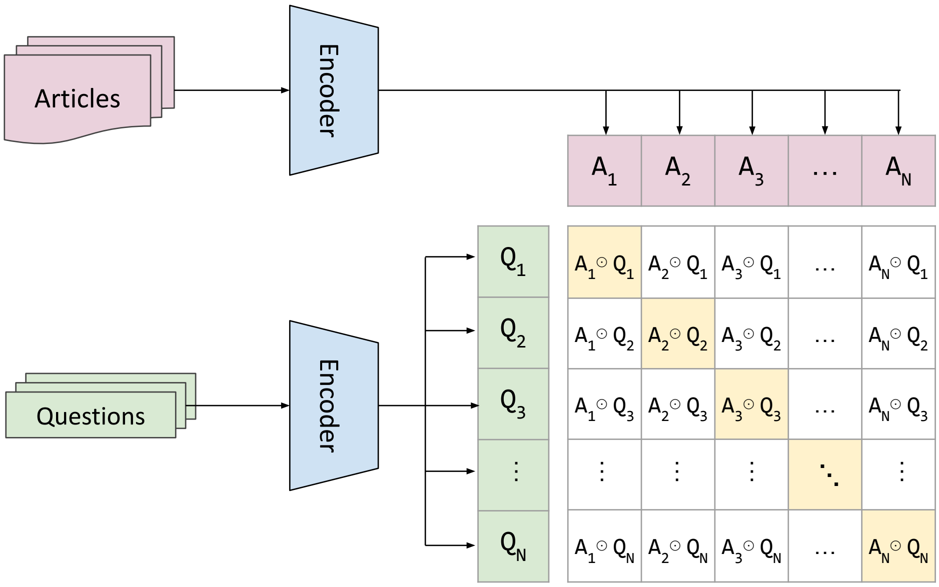

Figure 3: Standard Siamese Bi-Encoder setting with in-batch negatives, which we use for fine-tuning. Articles and Questions are encoded separately with the same model into vectors. For each question Q_i, the relevant article A_i is the positive sample, while all other articles in the batch are used as negatives. $\odot$ represents the cosine similarity operator.

In addition, we experiment with context-independent word embeddings, using word2vec (Mikolov et al., 2013b,a) for Dutch (Tulkens et al., 2016) and French (Fauconnier, 2015), as well as fastText (Bojanowski et al., 2017) for both Dutch and French (Grave et al., 2018). To construct embeddings of text chunks from these models, we apply mean-pooling to the word embeddings, with the exception of LaBSE, which uses the [CLS] token representation. In all cases, cosine similarity is employed to score similarity between the embeddings.

The evaluation is conducted on a single GPU with 48GB of RAM for E5$_{mistral-7b}$ and BGE-Multilingual-Gemma2. For other models, we use a single GPU with 8GB of RAM. Each experiment takes between five minutes for smaller models and up to 30 minutes for larger models.

4.1.3 Fine-tuned models

Foundation models can achieve competitive results compared to zero-shot retrieval models when fine-tuned on domain-specific data. In our evaluations, we select RobBERT-2023 (Delobelle and Remy, 2024) and Tik-to-Tok (Remy et al., 2023) for Dutch, and CamemBERT (Martin et al., 2020) and Flaubert (Le et al., 2020) for French. We also include XLM-Roberta to examine the potential advantage of language-specific models over the generic multilingual ones.

We primarily follow the experimental setup of BSARD and fine-tune the models in a Siamese setting (Reimers and Gurevych, 2019), which encodes the query and document via the same model (Figure 3). We optimize the contrastive loss with a temperature of 0.05 and in-batch negatives (Henderson et al., 2017; Karpukhin et al., 2020) with a batch size of 22. The optimization is performed using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. The learning rate undergoes a warm-up over the first 500 steps, followed by linear decay. Following Louis and Spanakis (2022), training is performed for 100 epochs, which takes 4.5-5.5 hours (depending on model size) on a single GPU with 24GB of RAM. Finally, we employ cosine similarity to score the embeddings.

4.2 Metrics

To assess the performance of our models, we employ standard 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).

5 Results and Discussion

In this section, we present the performance of various retrieval models evaluated on the Dutch and French subsets of the bBSARD dataset (see Tables

T	Model	Size	R@100	R@200	R@500	MAP@100	MRR@100	nDCG@10	nDCG@100
	TF-IDF	-	39.21	46.38	52.76	8.53	14.25	12.38	16.74
	BM25	-	40.19	47.95	54.57	16.07	22.63	20.07	23.57
	word2vec	-	41.06	51.02	58.94	8.28	15.27	11.66	17.05
	fastText	-	31.47	38.26	49.08	7.27	12.67	8.89	13.86
	mE5$_{small}$	118M	45.43	52.25	61.10	13.42	21.79	17.67	22.79
	mContriever	178M	47.92	58.38	68.32	11.38	20.15	14.83	21.82
	DPR-XM	277M	40.44	46.12	53.16	13.57	21.78	16.40	21.79
	mE5$_{base}$	278M	50.14	57.68	65.30	16.47	25.64	20.81	26.49
	mGTE	305M	52.78	61.97	73.09	15.86	24.92	20.08	26.80
	LaBSE	471M	20.51	28.42	42.18	2.34	6.60	3.50	7.18
	mE5$_{large}$	560M	58.35	65.83	70.83	21.88	34.28	28.47	33.51
	mE5$_{large-instruct}$	560M	59.48	66.80	75.21	18.66	29.93	24.84	31.33
	BGE-M3	568M	61.12	67.20	77.56	18.31	30.40	24.04	31.21
	jina-embeddings-v3	572M	60.70	67.92	77.37	18.59	31.21	24.70	31.58
	E5$_{mistral-7b}$	7B	68.35	73.91	82.82	30.24	43.26	37.70	43.02
	BGE-Mult.-Gemma2	9B	69.94	76.23	81.28	25.07	37.66	30.95	39.11
	voyage-3	-	73.08	79.37	85.67	32.81	46.38	40.06	46.21
	embedding-3-large	-	**75.70**	**80.22**	**88.24**	29.73	42.99	36.83	44.40
✓	Tik-to-Tok$_{base}$	116M	73.90	79.02	83.29	39.24	45.69	42.75	49.90
✓	RobBERT-2023$_{base}$	125M	75.08	79.33	83.40	**40.51**	**47.68**	**44.76**	**51.36**
✓	XLM-Roberta$_{base}$	279M	62.06	68.26	75.40	26.61	32.00	30.65	37.10

Table 1: Retrieval performance of different models on the Dutch subset of bBSARD (test set). Evaluations are zero-shot for the dense models, except for the last 3 models (check-marked) which are fine-tuned.

T	Model	Size	R@100	R@200	R@500	MAP@100	MRR@100	nDCG@10	nDCG@100
	TF-IDF	-	41.69	51.05	60.22	8.74	12.85	11.34	17.45
	BM25	-	51.81	56.95	65.51	17.02	26.02	21.54	27.52
	word2vec	-	49.93	62.29	71.11	13.45	21.45	17.32	23.62
	fastText	-	24.84	32.36	43.88	5.05	10.03	7.40	10.65
	mE5$_{small}$	118M	46.26	51.74	59.25	13.67	23.48	18.49	23.03
	mContriever	178M	46.01	56.62	68.42	12.94	21.56	17.59	22.94
	DPR-XM	277M	40.91	47.34	55.13	10.83	19.31	14.31	19.74
	mE5$_{base}$	278M	47.62	56.60	63.60	16.76	26.25	21.90	26.28
	mGTE	305M	57.54	66.57	77.02	19.40	30.14	24.13	31.02
	LaBSE	471M	21.62	32.86	46.66	2.74	7.00	4.17	7.67
	mE5$_{large}$	560M	55.30	62.83	69.85	21.54	34.27	28.06	32.68
	mE5$_{large-instruct}$	560M	60.99	68.34	76.75	19.77	32.60	26.52	32.44
	BGE-M3	568M	60.76	69.02	79.81	19.40	31.38	25.38	32.08
	jina-embeddings-v3	572M	64.05	71.67	78.76	20.51	34.52	27.09	34.19
	E5$_{mistral-7b}$	7B	69.41	74.53	84.06	27.43	40.22	34.82	41.07
	BGE-Mult.-Gemma2	9B	71.44	77.81	83.73	30.06	43.72	36.36	43.46
	voyage-3	-	77.71	**82.68**	**88.76**	38.78	**54.60**	45.96	52.51
	embedding-3-large	-	75.47	80.70	87.58	33.72	46.51	40.54	47.33
✓	CamemBERT$_{base}$	111M	77.10	80.63	86.37	39.08	46.99	44.25	50.95
✓	FlauBERT$_{base}$	138M	**78.15**	81.59	85.84	**42.11**	49.82	**46.69**	**53.48**
✓	XLM-Roberta$_{base}$	279M	63.31	70.70	77.76	30.57	37.84	34.90	40.82

Table 2: Retrieval performance of different models on the French subset of bBSARD (test set). Evaluations are zero-shot for the dense models, except for the last 3 models (check-marked) which are fine-tuned.

1 and 2). In addition, we directly compare model effectiveness between two languages leveraging the parallel nature of the dataset.

5.1 Dutch Subset

As Table 1 shows, BM25 proves to be a strong baseline for the Dutch subset, with zero-shot dense models only fully outperforming it starting from 300 million parameters.

In the zero-shot setting, we observe a consistent improvement in performance as the model size grows, with the exception of LaBSE, which shows relatively lower results. The small-sized models (below 200M parameters), mE5$_{small}$ and mContriever, outperform the context-independent models (i.e. word2vec and fastText), and while mContriever achieves higher recall (R@100, R@200, R@500), mE5$_{small}$ is better across all other metrics. mE5$_{small}$ even outperforms the larger DPR-XM model in almost all metrics.

In the next zero-shot category (around 300M parameters), mGTE significantly outperforms mE5$_{base}$ in recall (R@100, R@200, R@500), while doing marginally worse across other metrics. For models up to 1 billion parameters, BGE-M3 and jina-embeddings-v3 show comparable results and are the best performers in recall (R@100, R@200, R@500), but E5$_{large}$ demonstrates superior performance in MAP@100, MRR@100, and nDCG (@10, @100). Finally, the largest open models, E5$_{mistral-7b}$ and BGE-Multilingual-Gemma2, outperform all other open models by a large margin. However, they lag behind proprietary models, voyage-3 and embedding-3-large, which are the best performers for the zero-shot setup.

As the lower section of the table shows, the high performance of proprietary models can be matched or topped by fine-tuning small models. In particular, fine-tuned RobBERT-2023$_{base}$ outperforms these models in MAP, MRR and nDCG metrics. Additionally, language-specific models demonstrate a significant advantage over the multilingual XLM-Roberta.

5.2 French Subset

Table 2 shows the results for the French subset of bBSARD. We observe trends similar to Dutch, with BM25 remaining competitive with the zero-shot dense models up to 300 million parameters.

Similarly, we observe a steady increase in performance in the zero-shot setup as the average model size increases, with the exception of LaBSE. Interestingly, the context-independent model word2vec outperforms not only the sub-200M models mE5$_{small}$ and mContriever, but also the larger DPR-XM model, while beating mE5$_{base}$ in recall. In the 300M-parameter category, mGTE outperforms the larger mE5$_{large}$ model in recall (R@100, R@200, R@500), and competes with BGE-M3 in MAP@100, MRR@100, and nDCG (@10, @100). Among the models up to 1 billion parameters, jina-embeddings-v3 achieves the highest performance in recall (R@100, R@200), MRR@100, and nDCG@100, while BGE-M3 performs better in recall@500, and E5$_{large}$ demonstrates the best results in nDCG@10. The largest open models, E5$_{mistral-7b}$ and BGE-Multilingual-Gemma2, show superior performance over other open options. However, the proprietary models, voyage-3 and embedding-3-large, outperform them by a large margin, with voyage-3 showing the best overall performance. Finally, we see the competitive performance of small fine-tuned models, with FlauBERT$_{base}$ beating voyage-3 in 4 out of 7 metrics.

5.3 Cross-Language Comparison

As bBSARD is a parallel dataset, we can directly compare Tables 1 and 2 to gain deeper insights into performance discrepancies between the French and Dutch models.

On average, models show a higher performance on the French subset compared to the Dutch subset (see Table 4 in Appendix C). This is perhaps most notable in BM25 and word2vec, where French models outperform their Dutch counterparts by more than 10 recall points (the clear outlier is fastText, which performs significantly better on the Dutch subset.) In addition, mGTE, jina-embeddings-v3, and voyage-3 do significantly better on the French subset than the Dutch. Other models gain 2-3 additional recall points in Dutch and perform comparably well across other metrics for both languages, with the exception of E5$_{mistral-7b}$ and DPR-XM. These models show slightly lower recall (R@100, R@200, R@500) in Dutch but achieve higher scores in other metrics. Finally, the best fine-tuned performer in French, FlauBERT$_{base}$, outperforms the top performer in Dutch, RobBERT-2023$_{base}$, and XLM-Roberta gains 3-5 points higher results when trained and evaluated on French.

In addition to potential translation issues (see 3.2) which particularly affect lexical models, one intuitive hypothesis on the origin of this advantage concerns the significant difference in data availability between the two languages. For example, while the original RobBERT model was pre-trained on a 39 GB corpus (Delobelle et al., 2020), CamemBERT and FlauBERT used 138 GB and 71 GB of data, respectively[10] (Le et al., 2020). However, further analysis is required to determine the significance of this factor, as well as other contributing parameters.

6 Conclusions and Future Work

In this paper we presented bBSARD, the bilingual version of the BSARD dataset (Louis and Spanakis, 2022). To curate bBSARD, we scraped parallel Dutch and French articles from the online Justel database and translated the BSARD questions into Dutch. In addition to a parallel bilingual legal corpus, bBSARD offers a much-needed retrieval benchmark for the Dutch language, allowing for more accurate and reliable evaluation of Dutch retrieval models.

Based on our dataset, we conducted extensive benchmarking of the retrieval task (ranking passages by their relevance to a given query) for Dutch and French, both in zero-shot and fine-tuned scenarios. These experiments confirm the status of simple lexical methods like BM25 as strong baselines, the superiority of closed-source commercial models like Voyage and OpenAI in zero-shot setting, and the possibility of outperforming them via fine-tuning small language-specific models like RobBERT and FlauBERT. We also observed an overall advantage for French compared to Dutch, in both zero-shot and fine-tuning scenarios.

We hope that our work encourages and facilitates the development of better Dutch retrieval models in the legal domain, which are an essential part of popular LLM-based methods like RAG. In the future, we would like to first improve bBSARD's quality by manually checking/correcting all translated questions, and then expand our work beyond the legal domain by curating and providing a more comprehensive benchmark for the retrieval task in Dutch. Another interesting research avenue considers the cross-lingual training potentials offered by a bilingual parallel dataset. In our experiments, we observed that XLM-Roberta performs better in Dutch when finetuned for 50+50 epochs on French+Dutch data, compared to 100 epochs on the Dutch subset. This suggests the possibility of leveraging the bilingual structure for additional gains in performance, specially for the lower-resource language, but to examine and explore its real significance more experiments need to be conducted.

Limitations

We primarily inherit limitations from BSARD, as this dataset serves as the foundation for our work. The retrieval corpus is limited to the 32 Belgian codes from federal (Belgian) and Walloon law. As a result, bBSARD does not cover the whole of Belgian law, particularly omitting codes specific to Flanders. In addition, these limitations make the retrieval process incomplete as a part of relevant articles might be missing. Since we scraped the Belgian articles from around May 2021, bBSARD does not contain the updated version of the Belgian law.

Given these limitations, bBSARD is not intended for obtaining any comprehensive legal information or advise. Its primary purpose is to benchmark retrieval models and gain insights into the current state of the art. In accordance with the BSARD license (`cc-by-nc-sa-4.0`), we release our dataset under the same terms.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. We would like to thank Franci Haest for contributing to the data collection and to Luna De Bruyne for checking the quality of the translations. We also acknowledge the use of the GPT-4o model for assisting with error checking and proofreading of this paper.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. *Mining text data*, pages 77–128.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Fire 2019 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation*, pages 4–6.

Paheli Bhattacharya, Parth Mehta, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya,

[10] 282 GB and 270 GB before filtering/cleaning.

and Prasenjit Majumder. 2021. Fire 2020 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 1–3, New York, NY, USA. Association for Computing Machinery.

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.

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*, pages 71–80.

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

Pieter Delobelle and François Remy. 2024. Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 13:193–203.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

European Union. 2012. Charter of fundamental rights of the european union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT. [Accessed: 5 October 2024].

Jean-Philippe Fauconnier. 2015. French word embeddings.

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.

Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. 2021. Tweac: transformer with extendable qa agent classifiers. *arXiv preprint arXiv:2104.07081*.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and discussion of the competition on legal information, extraction/entailment (coliee) 2023. *The Review of Socionetwork Strategies*, 18(1):27–47.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. Coliee 2022 summary: methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 51–67. Springer.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024a. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. *arXiv preprint arXiv:2402.15059*.

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.

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

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

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.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

OpenAI. 2024. New embedding models and api updates. Accessed: 2024-10-31.

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish i would have loved this one, but i didn't–a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. Coliee 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*, pages 196–210. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

François Remy, Pieter Delobelle, Bettina Berendt, Kris Demuynck, and Thomas Demeester. 2023. Tik-to-tok: Translating language models one token at a time: An embedding initialization strategy for efficient language adaptation. *arXiv preprint arXiv:2310.03477*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Gerard Salton and Chung-Shu Yang. 1973. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.

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, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Zibing Que, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. Stard: A chinese statute retrieval dataset with real queries issued by non-professionals. *arXiv preprint arXiv:2406.15313*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

VoyageAI. 2024. Voyage 3. Accessed: 2024-10-31.

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. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

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

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv e-prints*, pages arXiv–1911.

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. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Overview of cail2018: Legal judgment prediction competition. *arXiv preprint arXiv:1810.05851*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

A Appendix: Scraping and Aligning the Articles

Table 3 shows a detailed summary of codes scraped from the **Justel** portal for bBSARD, compared to the original BSARD dataset. For alignment, we first leverage the article names/codes (e.g. Art. 14bis), and then use an automatic pipeline (consisting of a length comparison filter followed by ChatGPT queries) to spot the absent, misaligned, or non-aligned articles, which we then add and/or align manually. The alignment issues are mainly due to rare discrepancies in the way articles are registered in French and Dutch pages, or between French pages and BSARD dataset (for example 'Art. 14.2' vs. 'Art. 14/2').

B Appendix: Translating the Questions

To translate the questions into Dutch, we prompt `GPT-4o`[11] with the following instruction and context for each question (temperature = 0).

Prompt:`"You will be provided with a legal question and a related article from Belgian legislation. Your task is to translate the question from French to Dutch. The article serves solely as context to ensure the accuracy in legal understanding and terminology, so do not include any part of it in the translation. Return only the translation of the question without any additional information.`

```
<article>: {article} </article>
<question>: {question} </question>
question translated to Dutch:"
```

We also translate the 3 meta-fields available for each question in BSARD, i.e. `category`, `subcategory`, `extra_description` (although they are not used in the experiments). For this, we first refer to the `www.helderrecht.be` website (the Dutch version for `www.droitsquotidiens.be`), and extract the available corresponding categories and subcategories (35% of the total). We then use these translation pairs as examples to prompt `GPT-4o` to translate the rest of the phrases.

C Appendix: Comparison of French and Dutch Results

Table 4 shows the average retrieval performance for different model types on the French and Dutch subsets of bBSARD (test set).

[11] gpt-4o-2024-08-06

		BSARD		bBSARD	
Authority	**Code**	#Articles	#Relevant	#Articles	#Relevant
Federal	Judicial Code	2285	429	2283	429
	Code of Economic Law	2032	98	2032	98
	Civil Code	1961	568	1961	568
	Code of Workplace Welfare	1287	25	1287	25
	Code of Companies and Associations	1194	0	1193	0
	Code of Local Democracy and Decentralization	1159	3	1158	3
	Navigation Code	977	0	977	0
	Code of Criminal Instruction	719	155	719	155
	Penal Code	689	154	689	154
	Social Penal Code	307	23	307	23
	Forestry Code	261	0	261	0
	Railway Code	260	0	260	0
	Electoral Code	218	0	217	0
	The Constitution	208	5	208	5
	Code of Various Rights and Taxes	191	0	189	0
	Code of Private International Law	135	4	134	4
	Consular Code	100	0	100	0
	Rural Code	87	12	87	12
	Military Penal Code	66	1	0	0
	Code of Belgian Nationality	31	8	31	8
Regional	Walloon Code of Social Action and Health	3650	40	3643	40
	Walloon Code of the Environment	1270	22	1143	22
	Walloon Code of Territorial Development	796	0	795	0
	Walloon Public Service Code	597	0	597	0
	Walloon Code of Agriculture	461	0	461	0
	Brussels Spatial Planning Code	401	1	401	1
	Walloon Code of Basic and Secondary Education	310	0	310	0
	Walloon Code of Sustainable Housing	286	20	279	20
	Brussels Housing Code	279	44	279	44
	Brussels Code of Air, Climate and Energy Management	208	0	208	0
	Walloon Animal Welfare Code	108	0	108	0
	Brussels Municipal Electoral Code	100	0	100	0
	Total	22633	1612	22417	1611

Table 3: Distribution of codes in BSARD and bBSARD (this work). "Relevant" articles are meant with respect to the question set.

T	**Model Type**	**Lang.**	**R@100**	**R@200**	**R@500**	**MAP@100**	**MRR@100**	**nDCG@10**	**nDCG@100**
	Lexical	FR	**46.75**	**54.00**	**62.87**	**12.88**	**19.44**	**16.44**	**22.49**
		NL	39.70	47.17	53.67	12.30	18.44	16.23	20.16
	CI dense	FR	**37.39**	**47.33**	**57.50**	**9.25**	**15.74**	**12.36**	**17.14**
		NL	36.27	44.64	54.01	7.78	13.97	10.28	15.46
	CD dense	FR	**56.79**	**64.24**	**72.81**	**20.54**	**31.83**	**26.09**	**31.89**
		NL	56.00	63.02	71.51	19.17	29.79	24.35	30.52
✓	CD dense	FR	**72.85**	**77.64**	**83.32**	**37.25**	**44.88**	**41.95**	**48.42**
		NL	70.35	75.54	80.70	35.45	41.80	39.39	46.12

Table 4: Average retrieval performance per model type on bBSARD (test set). CI and CD refer to context-independent and context-dependent models, respectively. All dense models are evaluated in zero-shot setting, except for the lower section (check-marked) which are fine-tuned.