LLM-based Embedders for Prior Case Retrieval

Damith Premasiri, Tharindu Ranasinghe, Ruslan Mitkov
School of Computing and Communications, Lancaster University, UK
{d.dolamullage, t.ranasinghe, r.mitkov}@lancaster.ac.uk

Abstract

In common law systems, legal professionals such as lawyers and judges rely on precedents to build their arguments. As the volume of cases has grown massively over time, effectively retrieving prior cases has become essential. Prior case retrieval (PCR) is an information retrieval (IR) task that aims to automatically identify the most relevant court cases for a specific query from a large pool of potential candidates. While IR methods have seen several paradigm shifts over the last few years, the vast majority of PCR methods continue to rely on traditional IR methods, such as BM25. The state-of-the-art deep learning IR methods have not been successful in PCR due to two key challenges: ***i***. Lengthy legal text limitation; when using the powerful BERT-based transformer models, there is a limit of input text lengths, which inevitably requires to shorten the input via truncation or division with a loss of legal context information. ***ii***. Lack of legal training data; due to data privacy concerns, available PCR datasets are often limited in size, making it difficult to train deep learning-based models effectively. In this research, we address these challenges by leveraging LLM-based text embedders in PCR. LLM-based embedders support longer input lengths, and since we use them in an unsupervised manner, they do not require training data, addressing both challenges simultaneously. In this paper, we evaluate state-of-the-art LLM-based text embedders in **four** PCR benchmark datasets and show that they outperform BM25 and supervised transformer-based models.

1 Introduction

Information retrieval (IR) systems have progressed through several paradigm shifts in the last few decades (Zhu et al., 2023; Plum et al., 2024). Initial IR methods relied on term-based methods such as BM25 (Robertson et al., 2009) and boolean logic, focusing on keyword matching for document retrieval (Chowdhury, 2010). The paradigm gradually shifted with the introduction of vector space models, enabling a more sophisticated understanding of the semantic relationships between queries and documents (Salton et al., 1975). Initially, these models relied on statistical language models (Song and Croft, 1999), but in recent years, neural vector space models have achieved remarkable performance in IR (Guo et al., 2016; Xiong et al., 2021). More recently, large language models (LLMs) have been integrated into these vector space models as embedders, further improving the performance (Ma et al., 2024; Ranasinghe et al., 2025). LLM-based embedders have dominated text retrieval benchmarks such as MTEB (Muennighoff et al., 2023) and BEIR (Thakur et al., 2021).

Prior case retrieval (PCR) is an IR application where the goal is to retrieve cases from a large legal database of historical cases that are similar to a given query case (Fang et al., 2022; Feng et al., 2024; Li et al., 2021; Tran et al., 2020). PCR holds substantial practical value, regardless of the legal system a country follows. In the Common Law System (as in the United Kingdom and India), legal professionals, such as lawyers and judges, use precedents (previously decided court cases) to support their arguments and help achieve their desired outcome in the present case (Shulayeva et al., 2017). Even in the Civil Law System (like in China and France), where legal arguments primarily rely on statutes, PCR remains essential as it offers key reference details, including the relevant statutes for past cases and the court's rulings, serving both legal experts and those seeking legal advice (Li et al., 2024b).

While the general domain IR systems have progressed into neural retrieval models, PCR systems still rely largely on traditional and term-based methods such as BM25 (Robertson et al., 2009). Re-

Proceedings of Recent Advances in Natural Language Processing,pages 980–988
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-112

searchers participating in COLIEE have demonstrated that BM25 serves as a strong baseline, and most top-performing systems have employed models based on BM25 combined with other techniques, such as TF-IDF and XG-Boost (Joshi et al., 2023). For example, a traditional language modelling approach (Ponte and Croft, 1998) proposed in 1998 won the first place (Ma et al., 2021), and a vanilla BM25 got second place (Rosa et al., 2021) in COLIEE 2021 (Rabelo et al., 2022). Furthermore, many researchers such as Askari et al. (2021) and Joshi et al. (2023), show that BM25 outperforms many supervised transformer-based retrieval approaches.

The limited success of neural retrieval models in PCR can be attributed to two main reasons.

1. Long court cases - The court cases are lengthy in nature. Many state-of-the-art neural retrieval models, because of their reliance on BERT models (Devlin et al., 2019), have a context limit of 512 tokens (Khattab and Zaharia, 2020; Ren et al., 2021). While researchers have attempted to apply these neural retrieval models to PCR using techniques such as truncating court cases, they often result in information loss and suboptimal results (Askari et al., 2021; Joshi et al., 2023; Nguyen et al., 2021; Premasiri et al., 2023).

2. Lack of training data - The neural retrievals are supervised machine learning models and require a large number of training instances (Khattab and Zaharia, 2020; Ren et al., 2021). While several large PCR datasets exist, they are limited to a few languages and courts, and it is not always possible to find PCR datasets that are large enough to properly train neural retrievals, resulting in reduced performance.

The recently emerged LLM-based text embedders (Muennighoff, 2022; Lee et al., 2024) can simultaneously address both of these challenges in PCR. First, they take up to 32,000 tokens as input, which is under the length of most court cases, addressing the first challenge we discussed. Secondly, as we will discuss in Section 3.2, LLM-based text embedders can be utilised in an unsupervised manner in the IR tasks, eliminating the need for model training and addressing the second challenge (Ranasinghe et al., 2025). Furthermore, Ni et al. (2022) demonstrate that LLM-based encoders also exhibit superior generalisability, showing significant improvements not only in the specific targeted scenario but also across a range of general tasks outside the fine-tuned domain. However, previous studies have not evaluated LLM-based text embedders in PCR benchmarks. In this research, we address this gap by answering the following two research questions (RQs).

RQ1 - How do the state-of-the-art LLM-based text embedders perform in different PCR benchmarks?
RQ2 - How well does the model ranking in the MTEB (Muennighoff et al., 2023) benchmark generalise to PCR benchmarks?

Answering these questions, in this paper, (1) we provide the **first comprehensive evaluation** of state-of-the-art LLM-based text embedders on PCR. (2) We show that our simple adaptation of LLM-based embedders **outperforms BM25 and transformer-based methods** in PCR datasets across multiple languages and jurisdictions. We release the model code and evaluation scripts for the purpose of research usage via GitHub[1]

2 Related work

With the increasing volume of cases, there is a growing demand for automatic precedent retrieval systems to assist practitioners by providing prior cases relevant to the current case (El Jelali et al., 2015). Therefore, PCR has remained an active area of research in the IR community (Breuker et al., 2005; Leburu-Dingalo, 2024; Tang et al., 2024a,b). Several datasets have been released for the PCR task such as LeCaRDv2 (Li et al., 2024b), C3RD (Ye and Li, 2024) and MUSER (Li et al., 2023b) for Chinese courts, IL-PCR (Joshi et al., 2023) for Indian courts, GerDaLIR (Wrzalik and Krechel, 2021) and LePaRD (Mahari et al., 2024) for United State's courts. Recent shared tasks, such as COLIEE (Goebel et al., 2024a, 2023; Kim et al., 2022) and AILA (Parikh et al., 2021; Bhattacharya et al., 2019) have facilitated the development of many PCR datasets.

SAILER (Li et al., 2023a) introduced a structure-aware pre-trained model for the prior case retrieval task. Utilising encoder-decoder architecture, SAILER proposes a fact encoder, a reasoning decoder and a decision decoder, which are pre-trained on Chinese and United States case law. CaseLink (Tang et al., 2024b) introduces a different approach for PCR by creating a Global Case Graph. They utilise semantic and legal charge relationships in addition to the reference relationships to populate the

[1]Available at https://github.com/DamithDR/case-retrieval.git

case graph. However, evaluating similarity among case law is a complex task. DELTA (Li et al., 2025) proposed an encoder-based pre-trained model with structural word alignment to methodically align the relevant facts closer and the irrelevant ones distant. Moreover, PCR has also been studied at the paragraph level. T.y.s.s. et al. (2024) created a specific dataset for paragraph retrieval in the European Court of Human Rights (ECtHR) and performed zero-shot and fine-tuned experiments with multiple encoder models. BERT-PLI (Shao et al., 2020) fine-tunes a BERT model for the sentence pair classification task and utilises the semantic relationships to calculate the relevance prediction using an interaction map.

Given the importance of information retrieval in the legal domain, multiple research events have been organised in this area. Few notable events are LIRAI (De Luca et al., 2023) workshop focused on information retrieval systems generally in the legal domain, and particularly in case law COLIEE (Goebel et al., 2024b) shared-task focuses on case law retrieval systems. The results of the competition suggest that BM25-like retrieval models are still effective in the context of lengthy text, as in case law.

3 Methodology

3.1 Data

Considering the diversity of jurisdictions and languages, we selected four popular PCR datasets: `IL-PCR` (Joshi et al., 2023), `COLIEE-2022` (Kim et al., 2023), `MUSER` (Li et al., 2023c), and `IRLeD` (Mandal et al., 2017). We further summarise the details of the datasets in Table 1. We used the test set (both candidates and queries) of `IL-PCR`, test queries in the `COLIEE-2022` and the whole dataset of `MUSER` and `IRLeD`, as there were no separate splits.

3.2 Modelling

As we mentioned earlier, we utilise LLM-based embedders in our research to investigate prior case retrieval across different jurisdictions. We choose three embedding models which are among the top five in `MTEB` (Muennighoff et al., 2023) leaderboard[2] as of October 2024[3]. Namely they are; `BAAI/bge-en-icl`[4] (Li et al., 2024a), `Salesforce/SFR-Embedding-2_R`[5] (Meng et al., 2024) and `dunzhang/stella_en_1.5B_v5`[6]. We used Salesforce/SFR-Embedding-2_R and dunzhang/stella_en_1.5B_v5 in combination with SentenceTransformer Python package (Reimers and Gurevych, 2019) while `BAAI/bge-en-icl` with FlagEmbedding Python package.

Following our motivation to address the challenge of the lengthiness of prior cases, LLM-based embedders help us to obtain a high-dimensional vector representation of each case. Unlike transformer-based models, these models support a high context length. As the models can have a high context length, they can capture most of the information in lengthy case documents. The first step is to obtain embeddings for each candidate and query case for all datasets separately, utilising the above-mentioned models.

For the retrieval process, we iterate over each

Algorithm 1 Ranking Court Cases using Precomputed LLM Embeddings and MAP Evaluation for Multiple Datasets

Require: Dataset collections $\mathcal{D} = \{D_1, D_2, \ldots, D_5\}$ (court cases), Embedding model M
Require: Queries $Q = \{q_1, q_2, \ldots, q_m\}$ for each dataset in $\mathcal{D}$
Precompute embeddings e_q for each query $q \in Q$ using M
for each dataset $D \in \mathcal{D}$ **do**
  Precompute embeddings e_c for each candidate case $c \in D$ using M
  for each query $q \in Q$ **do**
    for each candidate case $c \in D$ **do**
      Calculate cosine similarity $s(q,c) = \frac{e_q \cdot e_c}{\|e_q\| \|e_c\|}$
      Append $(c, s(q,c))$ to list of candidates for q
    end for
    Sort candidates by descending similarity scores
    $R_q \leftarrow$ ranked list of candidates for q in dataset D
    for $k = 1$ to Top-K candidates in R_q **do**
      Calculate Precision@k, Recall@k, and F-score@k
    end for
  end for
  Calculate Mean Average Precision (MAP) across all queries in dataset D
end for

[2]Available at `https://huggingface.co/spaces/mteb/leaderboard`

[3]`NV-EMBED`(Lee et al., 2024) is the best model in the MTEB leaderboard as of October 2024. However, we were unable to use the model on an NVidia L40 48G GPU. Therefore, we only used the models ranked 2nd, 3rd and 4th in MTEB.

[4]Available at `https://huggingface.co/BAAI/bge-en-icl`

[5]Available at `https://huggingface.co/Salesforce/SFR-Embedding-2_R`

[6]Available at `https://huggingface.co/dunzhang/stella_en_1.5B_v5`

	IL-PCR (Joshi et al., 2023)	COLIEE 2022 (Kim et al., 2023)	MUSER (Li et al., 2023c)	IRLeD (Mandal et al., 2017)
Total No of queries	237	300	100	200
Total no of candidates	1727	1263	1038	2000
Average no of words in queries	6766.32	5107.03	1993.12	7801.15
Average no of words in candidates	7046.36	4700.66	1747.52	7294.31
Language	English	English	Chinese	English
Jurisdiction	India	Canada	China	India

Table 1: Details of the datasets

query case embedding and calculate the cosine similarity with all candidate embeddings. We rank the most similar embeddings with a higher rank and the least similar ones with a lower rank. We use cosine similarity as our primary metric for calculating the similarity. Our retrieval algorithm is shown in Algorithm 1.

We calculate the Mean Average Precision (MAP) as our primary evaluation metric. We also calculate precision@k, recall@k and F score@k values where k={1, 5, 10, 15...50 and 100} following the recent research in PCR (Joshi et al., 2023). We use two baselines. First, we employ BM25 (Robertson et al., 2009), which is a strong and popular baseline for PCR, as we mentioned before. As the second baseline, we employ sentence-transformer with LEGAL-BERT. We trained a LEGAL-BERT model using the training set of the IL-PCR dataset. We used the positive samples from the dataset and added five negative samples for each query case to create the training data. The model was trained with learning_rate=2e-5, epoch=1, batch_size=16. The resulting model was used to create embeddings following the same method as other models to evaluate them.

4 Results and Analysis

Table 2 summarises the MAP values and best F scores achieved by each model and respective k values. Figure 1 illustrates the F-score curves for all datasets using all models.

As can be seen in the results, LLM-based embedders outperform the BM25 baseline with a clear margin in all datasets, answering our **RQ1**. As can be seen, for all the k values, LLMs perform better than BM25, showing their effectiveness. IL-PCR dataset shows a 0.16 improvement in F score compared to BM25 in SFR-Embedding-2_R model while reporting 0.47 MAP score. In COLIEE dataset, both SFR-Embedding-2_R model stella_en_1.5B_v5 and show 0.06 improvement in MAP compared to BM25. Furthermore, in IL-PCR, all LLM-based encoders outperform the supervised legal-bert model. In other datasets, too, LLM-based embedders outperform the legal-bert model, showing that they generalise well compared to other unsupervised models. Overall, it is clear that LLM-based embedders provide a promising solution to PCR.

From the LLMs SFR-Embedding-2_R shows the best performance for three out of four datasets from both MAP and F scores, while stella_en_1.5B_v5 is the best performer for the MUSER dataset. However, it should be noted that bge-en-icl is the best model out of these models in the MTEB benchmark, yet it does not outperform other models in the PCR tasks. With this finding, we answer **RQ2**, the model ranking in the MTEB benchmark does not generalise into the PCR benchmarks. While MTEB benchmark contains IR tasks, it does not contain any PCR tasks, which explains our observation to **RQ2**.

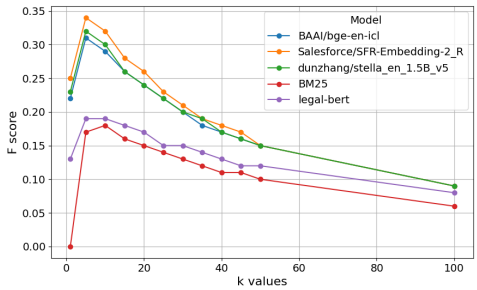
5 Conclusion

In this paper, we empirically showed that state-of-the-art LLM-based embedding models in MTEB benchmark outperform BM25 in multiple PCR datasets in multiple jurisdictions. However, as MTEB does not contain any PCR tasks, the model ranking in MTEB is not reflected in PCR datasets. Overall, LLM-based embedding models provided better results in all the PCR datasets, outperforming popular baselines, BM25 and other supervised baselines.

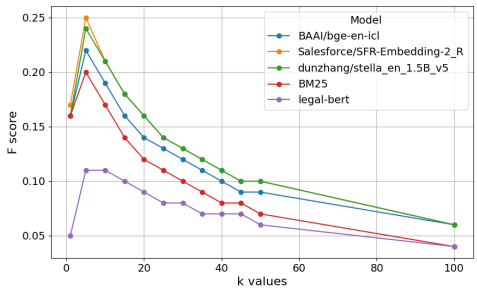
As the first comprehensive evaluation of LLM-based embedding models in the PCR task, this research will open several future research directions. First, the IR community needs to incorporate PCR datasets widely into IR benchmarks. Secondly, LLM-based embedders should be trained in PCR tasks so that they will provide better results than the unsupervised approach.

Model	IL-PCR			COLIEE			IRLeD			MUSER		
	MAP	F	k	MAP	F	k	MAP	F	k	MAP	F	k
bge-en-icl	0.42	0.31	5	0.29	0.22	5	0.25	0.25	5	0.10	0.08	25
SFR-Embedding-2_R	**0.47**	**0.34**	5	**0.32**	**0.25**	5	**0.27**	**0.27**	5	0.12	0.10	25
stella_en_1.5B_v5	0.44	0.32	5	0.32	0.24	5	0.26	0.25	5	**0.14**	**0.11**	25
LEGAL-BERT	0.27	0.19	5	0.14	0.11	5	0.09	0.08	10	0.04	0.04	10
BM25	0.16	0.18	10	0.26	0.20	5	0.20	0.20	5	0.12	0.10	25

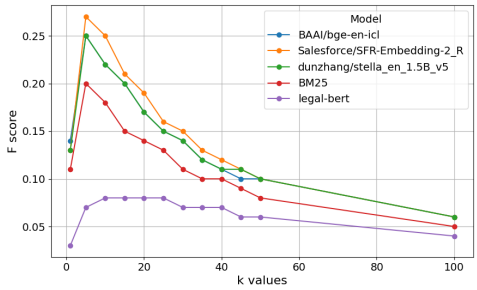
Table 2: Model performance on different datasets. Column Model shows the model used in the experiment and columns IL-PCR, COLIEE, IRLeD, MUSER show the dataset used for the experiment. Column MAP shows the mean average precision results, and column F score shows the best F score achieved by the model for the dataset. Column k value indicates the corresponding k value to the best F score. The only instance where we used a supervised model is coloured in blue.



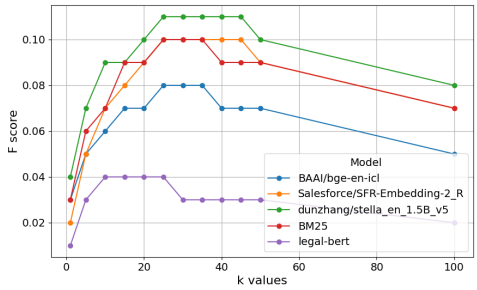
(a) IL-PCR



(b) COLIEE



(c) IRLeD



(d) MUSER

Figure 1: Change of F score with respect to k values.

Acknowledgements

We would like to thank the anonymous reviewers for their positive and valuable feedback. We further thank the creators of the datasets used in this paper for making the datasets publicly available for our research.

The experiments in this paper were conducted in UCREL-HEX (Vidler and Rayson, 2024). We would like to thank John Vidler for the continuous support and maintenance of the UCREL-HEX infrastructure, which enabled the efficient execution of our experiments.

Limitations

We acknowledge that there are bigger models with higher-dimensional representations; however, we do not conduct experiments on these models due to hardware resource limitations.

We used cosine similarity as our only relevancy metric; however, there are other methods that can be explored to get the similarity between two vectors. Furthermore, our experiments are currently limited to a few jurisdictions and languages, such as English and Chinese, which we plan to expand in the future, as well as developing specific models for PCR by further pre-training on PCR datasets.

References

AA Askari, SV Verberne, O Alonso, S Marchesin, M Najork, and G Silvello. 2021. Combining Lexical and Neural Retrieval with Longformer-Based Summarization for Effective Case Law retrieva. In *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES)*.

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.

Joost Breuker, André Valente, and Radboud Winkels. 2005. *Use and reuse of legal ontologies in knowledge engineering and information management*, page 36–64. Springer-Verlag, Berlin, Heidelberg.

Gobinda G Chowdhury. 2010. *Introduction to modern information retrieval*. Facet publishing.

Ernesto William De Luca, Manuel Fiorelli, Davide Picca, Armando Stellato, and Sabine Wehnert. 2023. Legal information retrieval meets artificial intelligence (lirai). In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Soufiane El Jelali, Elisabetta Fersini, and Enza Messina. 2015. Legal retrieval as support to emediation: matching disputant's case and court decisions. *Artif. Intell. Law*, 23(1):1–22.

Jingxin Fang, Xuwei Li, and Yiguang Liu. 2022. Low-resource similar case matching in legal domain. In *Artificial Neural Networks and Machine Learning – ICANN 2022*, pages 570–582, Cham. Springer Nature Switzerland.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 472–480.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024a. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 109–124. Springer.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024b. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *New Frontiers in Artificial Intelligence: JSAI International Symposium on Artificial Intelligence, JSAI-IsAI 2024, Hamamatsu, Japan, May 28–29, 2024, Proceedings*, page 109–124, Berlin, Heidelberg. Springer-Verlag.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. U-CREAT: Unsupervised case retrieval using events extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

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.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2023. Coliee 2022 summary: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12–17, 2022, Revised Selected Papers*, page 51–67, Berlin, Heidelberg. Springer-Verlag.

Tebo Leburu-Dingalo. 2024. Towards a framework for legal case retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3078, New York, NY, USA. Association for Computing Machinery.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

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

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. Sailer: Structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1035–1044, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, Qian Dong, and Yiqun Liu. 2025. Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27072–27080.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024b. 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.

Jieke Li, Min Yang, and Chengming Li. 2021. Clc-rs: A chinese legal case retrieval system with masked language ranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4734–4738, New York, NY, USA. Association for Computing Machinery.

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023b. Muser: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5336–5340.

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023c. Muser: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5336–5340, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, New York, NY, USA. Association for Computing Machinery.

Yixiao Ma, Yunqiu Shao, Bulou Liu, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Retrieving Legal Cases from a Large-scale Candidate Corpus. In *Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment (COLIEE)*.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex Pentland. 2024. LePaRD: A large-scale dataset of judicial citations to precedent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9863–9877, Bangkok, Thailand. Association for Computational Linguistics.

Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the fire 2017 irled track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

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

Ha-Thanh Nguyen, Phuong Minh Nguyen, Thi-Hai-Yen Vuong, Quan Minh Bui, Chau Minh Nguyen, Binh Tran Dang, Vu Tran, Minh Le Nguyen, and Ken Satoh. 2021. JNLP Team: Deep Learning Approaches for Legal Processing Tasks in COLIEE 2021. *arXiv preprint arXiv:2106.13405*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripa Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2021. Aila 2021: Shared task on artificial intelligence for legal assistance. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 12–15.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2024. Guided distant supervision for multilingual relation extraction data: Adapting to a new language. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 7982–7992, Torino, Italia. ELRA and ICCL.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 275–281, New York, NY, USA. Association for Computing Machinery.

Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2023. Can model fusing help transformers in long document classification? an empirical study. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 871–878, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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*.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025. MUSTS: MUltilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. 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.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, BM25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3501–3507. International Joint Conferences on Artificial Intelligence Organization. Main track.

Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126.

Fei Song and W. Bruce Croft. 1999. A general language model for information retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, page 316–321, New York, NY, USA. Association for Computing Machinery.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024a. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. In *Advances in Information Retrieval*, pages 80–95, Cham. Springer Nature Switzerland.

Yanran Tang, Ruihong Qiu, Hongzhi Yin, Xue Li, and Zi Huang. 2024b. Caselink: Inductive graph learning for legal case retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2199–2209, New York, NY, USA. Association for Computing Machinery.

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)*.

Vu Tran, Minh Le Nguyen, Satoshi Tojo, and Ken Satoh. 2020. Encoded summarization: summarizing documents into continuous vector space for legal case retrieval. *Artificial Intelligence and Law*, 28(4):441–467.

Santosh T.y.s.s., Elvin A. Quero Hernandez, and Matthias Grabmair. 2024. Query-driven relevant paragraph extraction from legal judgments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13442–13454, Torino, Italia. ELRA and ICCL.

John Vidler and Paul Rayson. 2024. UCREL - Hex; a shared, hybrid multiprocessor system. `https://github.com/UCREL/hex`.

Marco Wrzalik and Dirk Krechel. 2021. GerDaLIR: A German dataset for legal information retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Fuda Ye and Shuangyin Li. 2024. Milecut: A multi-view truncation framework for legal case retrieval. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1341–1349, New York, NY, USA. Association for Computing Machinery.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.