A Reproducibility Study of Graph-Based Legal Case Retrieval

Gregor Donabauer
Information Science
University of Regensburg
Regensburg, Germany
gregor.donabauer@ur.de

Udo Kruschwitz
Information Science
University of Regensburg
Regensburg, Germany
udo.kruschwitz@ur.de

Abstract

Legal retrieval is a widely studied area in Information Retrieval (IR) and a key task in this domain is retrieving relevant cases based on a given query case, often done by applying language models as encoders to model case similarity. Recently, Tang et al. proposed CaseLink, a novel graph-based method for legal case retrieval, which models both cases and legal charges as nodes in a network, with edges representing relationships such as references and shared semantics. This approach offers a new perspective on the task by capturing higher-order relationships of cases going beyond the stand-alone level of documents. However, while this shift in approaching legal case retrieval is a promising direction in an understudied area of graph-based legal IR, challenges in reproducing novel results have recently been highlighted, with multiple studies reporting difficulties in reproducing previous findings. Thus, in this work we reproduce CaseLink, a graph-based legal case retrieval method, to support future research in this area of IR. In particular, we aim to assess its reliability and generalizability by (i) first reproducing the original study setup and (ii) applying the approach to an additional dataset. We then build upon the original implementations by (iii) evaluating the approach's performance when using a more sophisticated graph data representation and (iv) using an open large language model (LLM) in the pipeline to address limitations that are known to result from using closed models accessed via an API. Our findings aim to improve the understanding of graph-based approaches in legal IR and contribute to improving reproducibility in the field. To achieve this, we share all our implementations and experimental artifacts with the community.[1]

CCS Concepts

• **Information systems → Information retrieval**.

Keywords

Legal Case Retrieval, Graph Neural Networks, Reproducibility.

ACM Reference Format:
Gregor Donabauer and Udo Kruschwitz. 2025. A Reproducibility Study of Graph-Based Legal Case Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730282

1 Introduction

Legal retrieval has attracted growing attention within the Information Retrieval (IR) community over time. This becomes evident from various events and projects related to the topic[2] [3] [10, 11, 15, 33].

One specific legal IR task, legal case retrieval, falls within the field of legal justice and focuses on retrieving relevant cases based on a given query case [8, 22, 23]. Legal experts can then analyze these cases to make informed judgments about the case in question [17].

Typically, case similarity in case retrieval is determined by using encoded representations of the individual case texts. This can be done by applying language models specifically trained on legal case data [17, 21], using large language models (LLMs) to preprocess the texts [28], or representing legal semantics within each case text as graphs, which are then aggregated using graph neural networks (GNNs) [29].

More recently, Tang et al. [30] proposed an approach, called CaseLink, that goes beyond the document level by considering naturally occurring relationships between cases in a network structure. They argue that both cases and legal charges can be represented as nodes in a graph, while relationships between them (such as references, semantics or higher-order relationships) can be modeled as edges. This contextual information, which is then processed with a GNN, has the potential to uncover relationships that are not visible when cases are viewed on an individual level, and offers a novel direction in the field of legal case retrieval.

While the potential benefits of integrating contextual and structured information into legal IR have repeatedly been highlighted [9, 18], there has been little research in this area so far.

Consequently, CaseLink introduces a novel perspective on legal case retrieval in an understudied area of graph-based legal IR. However, challenges related to the reproducibility of published work have been highlighted in recent years [4, 16, 19, 24, 26, 34]. To address this, we aim to contribute to the field by evaluating the reproducibility of recent research to determine whether it is reliable, referenceable, and extensible for the future.

In our work, we first reproduce the work on graph-based legal case retrieval of Tang et al. [30], presented at SIGIR 2024. We then run additional ablation studies to evaluate the generalizability of the results and assess whether extending CaseLink with other concepts of graph machine learning can positively impact the findings.

Main Contributions. Our key contributions can be summarized as follows:

[1] https://github.com/doGregor/caselink_reproducibility

[2] https://legalai2020.github.io/

[3] https://trec-legal.umiacs.umd.edu/

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730282

i. We conduct a reproducibility study of graph-based legal case retrieval (CaseLink) using two benchmark datasets from the COLIEE 2022 and 2023 competitions.
ii. We extend this study by including an additional dataset from COLIEE 2024 to assess whether performance remains consistent across new data.
iii. We evaluate the performance of heterogeneous graphs, which account for different node and edge types, and compare it with the homogeneous graphs used in the original work.
iv. We explore the impact of replacing GPT-3.5 (as deployed in the original study) with an open LLM in the processing pipeline.
v. Lastly, we provide all of our implementations and artifacts to support the reproducibility of all results presented.

The remainder of this paper is structured as follows: Section 2 provides background information on the importance of Legal IR in the broader area of Professional Search as well as on recent challenges related to reproducibility in IR. In Section 3, we then outline our research questions before describing the methodology behind the work we reproduce in Section 4. Sections 5 and 6 detail our experimental setup and findings. Finally, we discuss our results and draw conclusions in Section 7, followed by considering the paper's limitations (Section 8) and ethical considerations (Section 9).

2 Background

2.1 Legal Retrieval

A major part of information searching happens in the workplace, where professionals handle large amounts of information [32]. Legal retrieval is a specialized domain within professional search [20] and focuses on identifying and retrieving information essential for legal decision-making [3]. It is performed by legal professionals, such as lawyers, with the primary objective of gathering evidence to answer legal questions or support specific legal positions and arguments [3].

One key task in this domain is legal case retrieval, which involves identifying relevant cases based on a given query case [8, 22, 23]. Precedents play an important role in constructing legal arguments in common law systems [23] and are also essential in civil law systems, where drawing analogies between relevant prior cases is necessary to ensure justice [13].

Most research in IR has concentrated on the optimization and evaluation of ranking algorithms for web search instead of professional search, for example due to its greater commercial value [32]. This also applies to the legal domain and legal retrieval, which includes legal case retrieval as a high-recall scenario and primarily targets legal professionals with specialized knowledge in law [23].

In summary, IR research has focused less on professional search compared to web search. Therefore, it is important to also concentrate on research in professional domains such as the legal domain. Explainability and transparency have been identified as key future research directions in this field [32]. In line with this, we aim to contribute by conducting a reproducibility study on legal case retrieval, with the goal of improving the understanding of retrieval algorithms, their reliability, and their generalizability.

2.2 Reproducibility Issues in IR

Recent research on reproducibility in IR highlights two challenges: (1) difficulties in reproducing the experimental results reported in the original studies and (2) assessing the generalizability of approaches beyond the datasets used in the original work.

The challenge of reproducing reported results can be observed across various IR tasks. For example, in unsupervised query generation for re-ranking computational complexity was too high to rerun experiments with a solid hardware setup [19]. Similarly, in abstractive summarization with semantic graphs, performance fell below the original study's baselines despite following the same experimental configuration [16]. This also demonstrates that even papers published at high-quality venues may lack sufficient detail for successful reproduction [16]. Furthermore, studies introducing novel methods often claim significant improvements over baselines, while recent work has shown that this is not always the case: For example, in session-based recommendation, GNN models were found to perform worse than baselines in a reproducability experiment [24]. In some cases, while general trends can be confirmed, reported performance metrics still show notable differences, as it was observed for a balanced topic-aware sampling method for improving PLM-based rankers [34].

The second key issue, assessing the generalizability of approaches, has also been demonstrated in various contexts: For example, retrievability score calculation techniques produced different score distributions when applied across different datasets, which shows a limited robustness of these methods [26]. Another example is a multi-aspect dense retriever, which was evaluated on an additional dataset and performed worse than a weaker baseline [4]. To even better understand generalizability, this study and other studies further analyze alternative components in the experimental setup as well as their impact on robustness and performance.

In summary, two important reproducibility challenges in IR are (1) achieving the originally reported results and (2) assessing the generalizability of approaches across datasets and experimental setups. In this work, we address these challenges by first verifying whether the original results of CaseLink can be reproduced under the same conditions and then going beyond the original experimental setup to evaluate generalizability with additional data and alternative pipeline components.

3 Research Questions

In this paper, we formulate and address the following research questions:

- **RQ1:** *Are the results of CaseLink on COLIEE 2022 and 2023 reproducible?*

As an initial step, we reproduce (different team, same experimental setup[4]) the retrieval experiments from the original work on the two benchmark datasets from the COLIEE legal case retrieval challenges from 2022 [15] and 2023 [10].

We face difficulties when rerunning the experiments based on the provided code, requiring communication with the authors to get the approach to work. Our findings then show quite large differences

[4] This definition is in line with the *current* ACM terminology guidelines https://www.acm.org/publications/policies/artifact-review-and-badging-current

between our reproduced results and the numbers reported in the original work.

- **RQ2:** *Does CaseLink achieve similar performance on the more recently published COLIEE 2024 dataset?*

The results reported in the original paper show different performance outcomes between the two COLIEE datasets from 2022 and 2023. We thus extend this evaluation by including a third dataset, the one from COLIEE 2024, to assess the consistency of performance.

Our findings show that performance on COLIEE 2024 is higher than that of COLIEE 2023 but falls short of the results achieved on COLIEE 2022.

- **RQ3:** *How does modeling the case-charge network as a heterogeneous graph influence performance of CaseLink?*

While different document types with different relations are connected within the CaseLink graph, the original paper uses homogeneous graphs to represent the data, which do not account for these variations. We extend the setup by using heterogeneous graphs to explicitly represent the differences in node and edge types and compare the results with those from homogeneous graphs in the original work.

We find that, despite not accounting for the differences in node and edge types, homogeneous graphs result in better performance compared to heterogeneous graphs. However, these differences in performance are not significant in most cases.

- **RQ4:** *Does plugging in an open LLM into the CaseLink preprocessing pipeline change the overall performance?*

The complete pipeline for generating the initial CaseLink node representations is based on PromptCase [28] and CaseGNN [29]. As part of PromptCase, an LLM is used to generate summaries of the original cases. As previous work has highlighted reproducibility issues with models that are not publicly disclosed, we replace the GPT-3.5 model from the original work with an open LLM from the Llama family to assess whether this change results in comparable performance.

Our findings demonstrate that an open alternative shows to be a feasible solution and even results in consistently better performance, with the Llama-based CaseLink outperforming the GPT-3.5-based setup on two out of the three evaluated datasets.

4 Methodology

Before presenting our experiments and results, we will first describe the case retrieval task in more detail and outline the methodology behind CaseLink. Additionally, we will provide a brief overview of PromptCase and CaseGNN, which are parts of the CaseLink pipeline.

4.1 Legal Case Retrieval

The task's starting point is a set of cases, $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$. Given a query case $q \in \mathcal{D}$, the objective is to retrieve all cases from $\mathcal{D}$ that are relevant to q. This can be expressed as $\mathcal{D}^* = \{d_i^* | d_i^* \in \mathcal{D} \land \mathit{relevant}(d_i^*, q)\}$, where $\mathit{relevant}(d_i^*, q)$ indicates that d_i^* is relevant to the query case q. In the legal domain, relevant cases are those that can serve as precedents to be referenced by the query case.

4.2 CaseLink Graph

The approach first constructs a homogeneous graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes and $\mathcal{E}$ represents the set of edges connecting these nodes. The graph contains two types of nodes: nodes representing cases $d \in \mathcal{V}$ and nodes representing charges $c \in \mathcal{V}$. Case nodes correspond to all documents from the original set of cases $\mathcal{D}$, while charge nodes are derived from a list of federal court acts[5] the cases are related to. This set of legal charges can be represented as $C = \{c_1, c_2, \ldots, c_m\}$. Each node is assigned a feature vector x_v, obtained using encoders that generate embeddings for the respective case or charge texts. The encoder can be any model capable of encoding case or charge texts, such as BERT [5] or SAILER [17]. As in the original CaseLink paper we will use CaseGNN [29] for that task.

An edge between nodes v and u in the graph is denoted as $e_{vu} \in \mathcal{E}$. The graph includes three types of relationships: (1) *case-case* edges, (2) *charge-charge* edges, and (3) *case-charge* edges.

Case-case edges connect cases that are intrinsically related and link their nodes in the graph. Pairwise BM25 scores are calculated between cases, and the k most similar case pairs are added to the edge set. *Charge-charge* edges model the naturally occurring relationships between different charges as multiple charges can co-occur within the same case within a legal system. Charges whose embeddings have a similarity score (calculated via dot product or cosine similarity) above a given threshold δ are connected via edges which indicates a higher likelihood of co-occurrence in similar cases. Finally, *case-charge* edges link cases to the charges they contain. This can simply be determined by identifying the presence of a charge's name in the case text. All edges across these three relationships are part of the final set of edges $\mathcal{E}$.

4.3 CaseLink Learning

A GNN consisting of two successive graph attention layers (GAT) [31] is used to aggregate information within the graph, resulting in updated node representations $\forall v \in \mathcal{V} : h_v^l = \mathrm{GNN}(h_v^{l-1}, h_u^{l-1} : u \in \mathcal{N}(v))$, where l represents the layer number and $h_v^0 = x_v$. To increase the expressiveness of the node embeddings, they are concatenated with their initial representations after the convolution steps via a residual connection. The method for generating these initial representations will be discussed in Section 4.4.

The objective is to distinguish the relevant cases from a large collection of cases based on the given query. The overall loss function is denoted as:

$$l = l_{\mathrm{InfoNCE}} + \lambda \cdot l_{\mathrm{DegReg}} \tag{1}$$

where λ is the coefficient used to weigh the two losses. We will briefly touch on the two loss functions below, for more detail we refer to the original CaseLink paper [30].

In l_{InfoNCE} a contrastive learning loss is used to distinguish between the relevant and non-relevant cases for a given query. The goal is to bring true relevant cases closer while pushing false relevant cases further away. Positive samples are taken from the ground

[5] The benchmark datasets we use in our experiments are related to laws of the Federal Court of Canada, thus the charges are extracted from a list of the Federal Courts Act and Rules of Canada.

truth, while easy negative samples are either randomly sampled or selected as hard negatives based on BM25 relevance scores.

$$l_{\text{InfoNCE}} = -log \frac{e^{\frac{(s(h_q,h_{d^+}))}{\tau}}}{e^{\frac{(s(h_q,h_{d^+}))}{\tau}} + \sum_{i=1}^{n_e} e^{\frac{(s(h_q, h_{d_i^{easy-}}))}{\tau}} + \sum_{i=1}^{n_h} e^{\frac{(s(h_q, h_{d_i^{hard-}}))}{\tau}}} \tag{2}$$

where q is the query case, d^+ are relevant cases and d^{easy-} as well as d^{hard-} are easy and hard negative cases respectively. n_e and n_h represent the number of easy and hard negative case samples. s is a similarity metric, for example cosine similarity, and τ is the temperature coefficient.

To balance the contrastive objective, which provides limited signals for candidates within the entire set of cases, an additional degree regularization loss l_{DegReg} is introduced. This regularization makes sure that each candidate case is connected to only a small number of cases within the whole case set which aligns the model more closely with real-world requirements.

$$l_{\text{DegReg}} = \sum_{i=1}^{o} \sum_{j=1}^{n} (\hat{A}_{ij}) \tag{3}$$

where $\hat{A}_{ij}$ is the pseudo adjacency matrix and $\hat{A}_{ij} = cos(h_i, h_j)$ based on the updated node features h_i, h_j as obtained by applying the GNN to the original graph. n represents all cases in the case pool in total and o the number of cases in $\mathcal{D}$.

For inference the CaseLink model is applied to a graph that is constructed based on a test set of cases $\mathcal{D}_{test}$. Relevance scores $s(q, d)$ between a query case q and a candidate case d can then be calculated as:

$$s(q, d) = \cos(h_q, h_d) \tag{4}$$

where h_q and h_d are the representations of query case q and candidate case d from CaseLink. The highest scoring candidates are the ones that are retrieved.

4.4 PromptCase and CaseGNN

CaseLink uses the document representations from the baseline model CaseGNN as the initial node representations. CaseGNN relies on document representations generated by PromptCase. Thus, running these two baselines step-by-step is an important part of the whole CaseLink pipeline. For a better understanding of the pipeline and the baselines, we briefly describe both systems below. For a more detailed description of both methods we refer to the respective papers.

4.4.1 PromptCase. The goal of **PromptCase** [28] is to generate more expressive case representations rather than using the raw case text directly as input to an encoder. In particular, for each case, two additional condensed versions of the original text are constructed, referred to as *facts* and *issues*.

For the *fact* representation, the factual section of a case is provided to an LLM with the instruction to summarize it in 50 words. The *issue* representation is created by extracting all sentences from the case text in which specific terms are replaced with placeholders (these are already included in the dataset), then concatenating them into a new case representation.

The *facts*, *issues*, and original case texts are then embedded separately using the pre-trained language model SAILER [17]. Their concatenated embeddings result in a more sophisticated case representation compared to using the original case text embedding alone.

These representations can be used either for directly comparing case similarity for case retrieval or as the initial case representation for methods that build upon such embeddings, such as CaseGNN or CaseLink.

4.4.2 CaseGNN. The goal of **CaseGNN** is to transform the unstructured case texts into structured text-attributed graphs and aggregate these into expressive case representations. First, relation triplets are extracted from the *fact* and *issue* texts that were generated by PromptCase. The extracted triplets are then used to construct separate text graphs G_{fact} and G_{issue} for *facts* and *issues* respectively. The text attributes within the graph nodes are embedded using the SAILER language model [17].

Additionally, each graph includes a virtual node that spans all other nodes within a graph (and thus can be interpreted as an overall *fact* or *issue* representation). This virtual node is represented by the corresponding *fact* or *issue* embedding generated in PromptCase. A GNN consisting of two GAT layers [31] is then applied to aggregate information from the *fact* and *issue* graphs, producing a separate embedding for each of them. The two embeddings linked to the same case are concatenated to form a comprehensive overall representation. The learning objective is the same as presented in Equation 2.

Again, these representations can be used either for directly comparing case similarity during case retrieval or as the initial case representation in CaseLink.

5 Experimental Setup

Datasets. The original experiments were conducted using two benchmark datasets from the COLIEE 2022 [15] and COLIEE 2023 [10] legal case retrieval competitions. These datasets consist of cases collected from the Federal Court of Canada, and the training and test sets for each dataset are disconnected from each other with no overlap. Since the cases are from Canada, Tang et al. used a list of Canadian legal charges[6] in their model to create the set of charge nodes. In addition to these two datasets, which were also used in the original paper, we include the COLIEE 2024 dataset [11], which follows the same structure. We report the key statistics of the three datasets in Table 1. We note that our token counts differ from those reported in the original paper. We used the NLTK package for tokenization to calculate these statistics as the original paper did not specify the method used for their calculations. However, the choice of tokenizer does not affect the actual data processing involved in reproducing the original work; it is used solely for computing the statistics presented in Table 1 for comparison purposes.

Metrics. In line with the original work we use the following standard information retrieval metrics to evaluate model performance:

[6] https://www.fct-cf.gc.ca/en/pages/law-and-practice/acts-and-rules/federal-court/

Dataset	COLIEE 2022		COLIEE 2023		COLIEE 2024	
	train	test	train	test	train	test
# Queries	898	300	959	319	1278	400
# Candidates	4415	1563	4400	1335	5616	1734
# Avg. relevant cases	4.68	4.21	4.68	2.69	4.16	3.91
Avg. case length (# tokens)	5609.64	5803.21	5457.58	4855.25	5322.20	5882.63
Largest case length (# tokens)	107772	72114	107772	52137	107772	125233

Table 1: Dataset statistics.

Precision (P), Recall (R), Micro F1 Score (Mi-F1), Macro F1 Score (Ma-F1), Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and normalized discounted cumulative gain (NDCG). We report results at $k = 5$ and use the same evaluation implementations as in the original CaseLink paper.

Statistical Testing. To compare performance across different runs, we conduct statistical tests. First, we divide the test collection into five equal subsets and compute the relevant metrics at the subset level. We then apply paired t-tests with Bonferroni correction at $p < 0.05$ to compare runs of multiple systems. Our evaluations are based on the NDCG@5 metric, which was used as the early stopping criterion during model training in the original study[7].

Large Language Models. As mentioned, running CaseLink requires to first run PromptCase and CaseGNN. The LLM used in PromptCase to generate summaries of case documents is OpenAI's GPT-3.5-Turbo, which was trained for instruction-following tasks. While the exact number of parameters for this model has not been publicly disclosed, a (now-withdrawn) paper by Microsoft suggests it may have around 20 billion parameters [25]. We also use Llama 3.1 [1] from Meta AI as an open[8] alternative which is available on huggingface[9]. We again use the version of the model that was trained for instruction-following with 8 billion parameters (Llama-3.1-8B-Instruct). We will provide more motivation for selecting this specific model for our experiments in Section 6.4.

Hyperparameters. For all experiments, we use the hyperparameters that are reported to result in the best performance according to the original paper. We will briefly describe them below and they are also all available on our Github.

PromptCase does not require setting any hyperparameter. In general, we follow the exact experimental setup from the original paper, for example for pre-processing.

For training the **CaseGNN** model, we use a learning rate of 0.000005 with a weight decay of 0.00005. We train for 1000 epochs with a batch size of 32. As mentioned in Section 4.4.2, training requires using a temperature parameter and (hard) negative samples for contrastive learning. We set $\tau = 0.1$ and the sample numbers to $n_e = 1$ and $n_h = 5$.

To construct the **CaseLink** graphs, we set $k = 5$ (number of BM25-based top similar case pairs) to get the *case-case* edges and $\delta = 0.9$ (embedding similarity threshold) to get the *charge-charge* edges. During model training, we set $\lambda = 0.001$ (balance between the two loss functions) and the number of negative samples d^{-} to $n_e = 1$ and $n_h = 5$. We again set the temperature parameter $\tau = 0.1$. We train with a batch size of 128 and a learning rate of 0.00001 (without weight decay) for 1000 epochs. The number of GNN layers in the neural network is set to 2.

6 Experiments and Results

In this section we describe our experiments and results to the research questions proposed in Section 1. Each subsection corresponds to one of the research questions. All experiments were executed using two Nvidia RTX A6000 GPUs with 48 GB VRAM each. For comparison between runs during discussion of our findings, we primarily use the NDCG metric, as Tang et al. select the reported results based on this measure.

6.1 Legal Case Retrieval Reproducibility

Our first goal is to find out whether the CaseLink results reported on the COLIEE 2022 and COLIEE 2023 datasets are reproducible. This is addressed by our first research question:

- **RQ1:** *Are the results of CaseLink on COLIEE 2022 and 2023 reproducible?*

To answer this, we reran all the steps required to get CaseLink results on these two datasets. This also includes running both PromptCase and CaseGNN, whose results we will report as strong baselines along BM25 as IR standard baseline. Despite the authors providing their implementations[10], we encountered several issues while reproducing the experiments, which we will outline below.

First, the code of the three approaches is located in three different repositories. This means that we could not simply run all steps that are needed for reproducing CaseLink based on the CaseLink repository. To resolve this, we merged the methods into one shared repository with a consistent folder structure which simplifies reproducibility and makes it easier to apply the approach to new datasets.

We were also facing problems when attempting to rerun the summary generation step in PromptCase, which in the original paper is based on GPT-3.5-Turbo. Since OpenAI's API model names change over time, we contacted the authors to clarify which specific model version they used during their experiments. Communication with the authors was quick and helpful, and they informed us about having used GPT-3.5-Turbo between June and December 2023. The model active during most of that time (*gpt-3.5-turbo-0613*, released on 13th of June 2023) is deprecated since June 2024. We therefore

[7] For implementation details of the original paper, see https://github.com/yanran-tang/CaseLink

[8] We refer to the model as *open* instead of *open-source* as, while its weights are publicly available, its license does not meet open-source standards.

[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[10] https://github.com/yanran-tang/CaseLink

used the closest model, *gpt-3.5-turbo-1106*, which was released on 6th of November 2023. This shows that using commercial LLMs is not fully reproducible, as support can be discontinued by the providing companies.

Next, we encountered issues when reproducing CaseGNN (the next step of the CaseLink pipeline after PromptCase). CaseGNN requires to (1) filter case data by year and (2) identify the top 50 matching cases based on BM25. However, scripts to run these two steps were initially not provided in the respective repository. After contacting the authors, they referred us to a script from another repository that could be adapted for year filtering, which resulted in the same results as in processed files provided in their GitHub repository. We added this script to our reproducibility repository to support future experiments. However, the instructions for generating the top 50 BM25 matches were vague, and despite testing various n-gram configurations, we were unable to reproduce their exact ranking files. We instead use the most similar results we could generate. The authors noted that BM25 can have some variations, which may have influenced the discrepancies in our results.

Additionally, we found minor errors related to incorrect file and folder path names in the provided scripts, which we had to correct to ensure the pipeline ran correctly. The final step before running CaseLink – saving the CaseGNN embeddings to be used as initial node representations for CaseLink – was missing and we implemented a solution by ourselves.

Results. We present the reproduced results for CaseLink along with the baselines BM25, PromptCase, and CaseGNN in Table 2 for the COLIEE 2022 dataset and Table 3 for the COLIEE 2023 dataset. In addition, we include a comparison of the absolute differences relative to the numbers reported in the respective original papers.

One surprising observation is that while we get nearly identical results for BM25 in terms of Precision, Recall, and F1 scores, the MRR, MAP, and NDCG scores were much higher on both datasets compared to those originally reported.

For the other methods (all related to CaseLink), the results are more diverse. On the COLIEE 2022 data PromptCase and CaseGNN both result in quite similar numbers as originally reported while on the COLIEE 2023 dataset the differences between reproduced and original results are quite high for both methods: For PromptCase performance is 1.6 (Precision) to 5.4 (MAP) lower, for CaseGNN the differences range from −4.9 (Precision) to −10.3 (NDCG).

Finally for CaseLink, which is the main interest of our reproducibility experiments, we observe a drop between 5.4 (NDCG) and 7.2 (Recall) on the COLIEE 2022 dataset and a drop of 3.8 (Precision) to 11.7 (MRR) on the COLIEE 2023 data compared to the numbers reported in the original paper.

Since PromptCase and CaseGNN are preprocessing components of CaseLink, the lower performance of CaseLink could be partially explained by the already lower performance of these two methods within the pipeline. However, on the COLIEE 2022 dataset, only CaseLink underperforms compared to the original results, while PromptCase and CaseGNN show performance that is very similar with the originally reported numbers.

Answer to RQ1. To summarize the findings related to our first research question, we were unable to reproduce the originally reported results of CaseLink on the COLIEE 2022 and COLIEE 2023 datasets. Across both datasets, our results were consistently lower for all metrics compared to those reported in the original paper.

The most surprising observations are the higher MRR, MAP, and NDCG scores for BM25. BM25 is known to be stable for binary relevance metrics but more sensitive to rank-based evaluations. Since Tang et al. only provided implementations for calculating the similarity matrix of cases without a full BM25 retrieval pipeline, we used our own implementations for that step. This may explain the high variation observed in rank-based scores.

The second observation of CaseLink performing worse on COLIEE 2023 than reported in the original paper could potentially be attributed to the already lower performance of PromptCase and CaseGNN which makes a further drop unsurprising. However, on COLIEE 2022, where PromptCase and CaseGNN perform stable, CaseLink still underperforms. One possible explanation is that CaseLink's graph structure is constructed across the entire corpus, and even small variations in edges could negatively impact its performance.

6.2 Performance on New Data

As observed in the results for COLIEE 2022 and 2023, CaseLink's performance variations between the two datasets were considerable (i.e. NDCG 64.9 vs. 38.6). To further assess CaseLink's generalizability, we evaluate the approach on the most recent COLIEE dataset from 2024 and want to address our second research question:

- **RQ2:** *Does CaseLink achieve similar performance on the more recently published COLIEE 2024 dataset?*

The COLIEE 2024 dataset [11], like those from COLIEE 2022 and 2023, is part of the same legal case retrieval competition and was released during the most recent edition. It is slightly larger than the previous datasets, as shown in the key statistics in Table 1.

As mentioned earlier, one issue of the original implementations was that they did not support datasets beyond those used in the original paper. This for example was evident in the use of hardcoded folder names within the code. To improve the applicability of the approach on new datasets, we generalized the folder structure so that it does not include any dataset-specific subnames anymore. Additionally, we revised all file paths in the scripts to align with the updated structure and implemented automatic saving of files in a unified location.

Results. Our results on the COLIEE 2024 dataset are presented in Table 4. CaseLink achieves the highest scores across all metrics except for MAP, which is different to COLIEE 2022 and 2023, where CaseGNN and BM25 respectively resulted in the highest scores for multiple metrics. Looking at NDCG score differences, for COLIEE 2022, CaseLink performed 4.4 lower than the best baseline CaseGNN (differences are not significant), and for COLIEE 2023, it dropped by 5.4 compared to BM25 (again, no significant difference).

In terms of absolute values, the NDCG score for COLIEE 2024 is 47.1, compared to 64.9 (+17.8) for COLIEE 2022 and 38.6 (−8.5) for COLIEE 2023. This places the performance on the new dataset between the results we got for the two previous datasets.

Answer to RQ2. To conclude, we answer the second research question with yes, CaseLink's performance on the COLIEE 2024 dataset is comparable to previous results, as the observed outcomes fall

Method	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
BM25	17.7	21.0	19.2	21.2	39.8	38.7	43.1*
diff. to BM25 in [30]	(−0.2)	(−0.2)	(−0.2)	(−0.2)	(+16.2)	(+13.3)	(+9.5)
PromptCase	17.5	20.8	19.0	21.0	34.2	33.0	37.9*
diff. to [28]	(+0.4)	(+0.5)	(+0.5)	(+0.5)	(−0.9)	(−0.9)	(−0.8)
CaseGNN	**32.9**	**39.0**	**35.7**	**40.5**	**66.0**	**63.6**	**69.3**
diff. to [29]	(−2.6)	(−3.1)	(−2.7)	(−1.9)	(−0.8)	(−0.8)	(±0.0)
CaseLink	30.9	36.7	33.5	37.5	61.1	58.4	64.9
diff. to [30]	(−6.1)	(−7.2)	(−6.6)	(−6.7)	(−6.2)	(−6.6)	(−5.4)

Table 2: Results on the COLIEE 2022 dataset. Significant differences on the NDCG@5 metric compared to CaseLink using paired t-tests and Bonferroni correction at $p < 0.05$ are marked with *. Bold indicates best result for this metric.

Method	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
BM25	16.5	30.6	21.4	22.1	**41.0**	**40.3**	**44.0**
diff. to BM25 in [30]	(±0.0)	(±0.0)	(±0.0)	(−0.1)	(+17.9)	(+19.9)	(+20.3)
PromptCase	14.4	26.8	18.8	19.3	27.4	26.6	31.0
diff. to [28]	(−1.6)	(−2.9)	(−2.0)	(−2.2)	(−5.3)	(−5.4)	(−5.2)
CaseGNN	12.8	23.8	16.6	16.7	29.2	28.3	32.5
diff. to [29]	(−4.9)	(−9.0)	(−6.4)	(−6.9)	(−9.7)	(−9.4)	(−10.3)
CaseLink	**17.1**	**31.8**	**22.3**	**22.7**	34.1	33.3	38.6
diff. to [30]	(−3.8)	(−6.6)	(−4.8)	(−5.5)	(−11.7)	(−11.0)	(−11.2)

Table 3: Results on the COLIEE 2023 dataset. Significant differences on the NDCG@5 metric compared to CaseLink using paired t-tests and Bonferroni correction at $p < 0.05$ are marked with *. Bold indicates best result for this metric.

Method	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
BM25	18.5	23.7	20.8	22.0	**42.9**	**41.4**	45.7
PromptCase	18.3	23.4	20.6	22.6	30.8	30.0	34.6*
CaseGNN	16.4	21.0	18.4	19.4	35.1	33.6	38.2*
CaseLink	**22.2**	**28.4**	**24.9**	**26.5**	**42.9**	41.2	**47.1**

Table 4: Results on the COLIEE 2024 dataset. Significant differences on the NDCG@5 metric compared to CaseLink using paired t-tests and Bonferroni correction at $p < 0.05$ are marked with *. Bold indicates best result for this metric.

between the performance levels achieved on the COLIEE 2022 and 2023 datasets.

The overview papers of COLIEE 2022 [15], 2023 [10], and 2024 [11] show that performance of submitted approaches to the legal retrieval challenges varies across years. This suggests that the datasets differ in complexity and teams submitting similar approaches are also facing higher/lower performance from year to year. Contributing factors may include the number of cases in each dataset and the distribution of relevant versus non-relevant pairs (see Table 1). Additionally, things such as the complexity of the legal language or how easy it is to construct meaningful case-charge networks from the case texts using the methods in CaseLink may play a role.

6.3 CaseLink with Heterogeneous Graphs

In the original CaseLink paper Tang et al. model the data as homogeneous graphs despite representing different types of nodes (case and charge) and edges (*case-case*, *charge-charge*, and *case-charge*). Recent research in other IR-related areas has shown that using heterogeneous graphs that explicitly account for differences in nodes and edges can improve performance, e.g. in fake news detection [6] or the medical domain [7]. Additionally, in a section called *Graph Extensions* Tang et al. mention positive implications that could result from modeling the data as a heterogeneous graph.

We thus construct a heterogeneous case-charge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that consists of a set of disjoint vertex sets $\mathcal{V} = \mathcal{V}_D \cup \mathcal{V}_C$ where $\mathcal{V}_D \cap \mathcal{V}_C = \emptyset$ as well as edges that are satisfying constraints according to the node types they link together. More specifically, we use three explicit types of edges: *case-case* ($(v_d, \tau_{DD}, u_d) \in \mathcal{E} \rightarrow v_d \in \mathcal{V}_D, u_d \in \mathcal{V}_D$), *charge-charge* ($(v_c, \tau_{CC}, u_c) \in \mathcal{E} \rightarrow v_c \in \mathcal{V}_C, u_c \in \mathcal{V}_C$) and *case-charge* ($(v_d, \tau_{DC}, v_c) \in \mathcal{E} \rightarrow v_d \in \mathcal{V}_D, v_c \in \mathcal{V}_C$). The graph structure itself as well as the training procedure are the same as in the original paper.

We apply a GNN with Heterogeneous Graph Transformer (HGT) [14] layers to aggregate the information in the heterogeneous graph and want to answer our third research question:

- **RQ3:** *How does modeling the case-charge network as a heterogeneous graph influence performance of CaseLink?*

Results. We present the results with heterogeneous graphs in Table 5. As observable, all of the metrics result in the highest scores with

Dataset	Graph Type	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
COLIEE 2022	homogeneous	**30.9**	**36.7**	**33.5**	**37.5**	**61.1**	**58.4**	**64.9**
	heterogeneous	28.3	33.6	30.7	34.9	60.2	**58.4**	63.9
COLIEE 2023	homogeneous	**17.1**	**31.8**	**22.3**	**22.7**	**34.1**	**33.3**	**38.6**
	heterogeneous	14.9	27.6	19.3	19.5	31.3	30.9	35.4
COLIEE 2024	homogeneous	**22.2**	**28.4**	**24.9**	**26.5**	**42.9**	**41.2**	**47.1***
	heterogeneous	19.8	25.4	22.2	23.2	39.4	37.4	42.9*

Table 5: CaseLink results with heterogeneous graphs. Significant differences on the NDCG@5 metric between the homogeneous and heterogeneous setup for each year using paired t-tests at $p < 0.05$ are marked with *. Bold indicates best result for this metric within the same dataset.

homogeneous graphs. However, for COLIEE 2022 and 2023, the differences between homogeneous and heterogeneous graphs are not statistically significant, while for COLIEE 2024, the homogeneous setting outperforms the heterogeneous one. The absolute differences for NDCG on heterogeneous graphs amount to −1.0 (COLIEE 2022), −3.2 (COLIEE 2023), and −4.2 (COLIEE 2024).

Answer to RQ3. In summary, our findings suggest that heterogeneous graphs do not result in advantages for CaseLink, despite prior research indicating so. However, the overall differences between the two setups are not significant on two out of three datasets.

This suggests that homogeneous graph embeddings are *good enough* (or even better) to model similarity of cases within the case-charge network. For example, if the important signals for modeling case similarity are primarily embedded in case-case relationships, then representing all entities as the same node type might be enough. Moreover, introducing heterogeneous node and edge types increases the complexity of the learned GNN model. This can lead to overfitting on underrepresented edge types or to the underutilization of certain node or edge types if they have little additional discriminative power.

6.4 Llama as LLM for Preprocessing

As mentioned in Section 6.1, using a commercial LLM as part of PromptCase resulted in challenges for reproducibility. Previous studies have also highlighted that many models that are not publicly disclosed and offer interaction only via API hinder reproducibility [2, 27]. We therefore want to evaluate how using an open alternative performs in context of CaseLink which is why we adopt Llama-3.1-8B-Instruct for the same task as outlined earlier (Section 5).

We select Llama-3.1-8B-Instruct for two main reasons: (1) This model, along with the entire Llama-3.1 family, was released in July 2023[11], which is in line with the release dates of the GPT-3.5-Turbo versions that were used in the original experiments; (2) As mentioned previously, OpenAI's GPT-3.5-Turbo is estimated to have around 20 billion parameters. The Llama-3.1 family includes models of varying sizes – 8 billion, 70 billion, and 405 billion parameters. Among these, the 8 billion parameter model we adapt in our experiments is the closest in scale to the used GPT-3.5 models. Llama-3.1-8B-Instruct thus serves as a comparable open alternative to the original GPT-3.5-Turbo model.

The final research question we want to answer is:

- **RQ4:** *Does plugging in an open LLM into the CaseLink preprocessing pipeline change the overall performance?*

Results. The results using Llama-3.1 in place of GPT-3.5-Turbo across all three COLIEE datasets are presented in Table 6. As we can see, the scores are consistently higher across all metrics and datasets when using the open alternative compared to the commercial LLM.

Although most differences are not statistically significant, we find that CaseLink, when combined with Llama, outperforms the GPT-3.5-based setup on the COLIEE 2023 and 2024 datasets. Similarly, CaseGNN achieves significantly better performance on the COLIEE 2024 dataset with Llama than with GPT-3.5.

Answer to RQ4. In response to our final research question, we find that using Llama-3.1-Instruct-8B as an open alternative to GPT-3.5-Turbo does indeed affect the overall performance of CaseLink. The impact is positive, as the Llama-based setup consistently results in higher scores across all evaluation metrics, for CaseLink even in significant improvements on two out of three datasets.

Interestingly, recent reproducibility studies in the field of IR have also identified Llama-3.1-Instruct-8B as one of the most stable LLMs evaluated [12]. In line with our findings, this suggests that Llama models generally deliver robust performance across a range of downstream tasks, potentially due to an optimized training strategy or improved architectural design.

7 Discussion and Conclusion

In this reproducibility study, we evaluated a recent legal case retrieval method called CaseLink, which models contextual information between cases and charges in graphs. We encountered a number of issues when reproducing the results reported in the original experiments, which is in line with findings from previous reproducibility tracks, where many papers reported difficulties in achieving the same results as the original studies they reproduced [4, 16, 19, 24, 26, 34].

Additionally, we conducted ablation studies on the new COLIEE 2024 dataset to assess the generalizability of the approach. The results are between those reported on datasets from previous COLIEE competitions, suggesting a consistent but varying performance across different data. We also explored alternative graph data setups, comparing heterogeneous networks with the homogeneous graphs used in the original work. However, these more complex

[11] Compare model release date on https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct or https://ai.meta.com/blog/meta-llama-3-1/

Dataset	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
COLIEE 2022							
PromptCase with Llama	**18.7**	**22.3**	**20.3**	**22.0**	**37.0**	**35.2**	**40.4**
with ChatGPT 3.5	17.5	20.8	19.0	21.0	34.2	33.0	37.9
CaseGNN with Llama	**36.0**	**42.8**	**39.1**	**44.0**	**68.5**	**66.1**	**71.1**
with ChatGPT 3.5	32.9	39.0	35.7	40.5	66.0	63.6	69.3
CaseLink with Llama	**31.5**	**37.5**	**34.2**	**38.3**	**63.0**	**61.4**	**67.3**
with ChatGPT 3.5	30.9	36.7	33.5	37.5	61.1	58.4	64.9
COLIEE 2023							
PromptCase with Llama	**15.4**	**28.5**	**20.0**	**20.6**	**29.3**	**28.7**	**32.7**
with ChatGPT 3.5	14.4	26.8	18.8	19.3	27.4	26.6	31.0
CaseGNN with Llama	**14.2**	**26.3**	**18.4**	**18.7**	**32.1**	**30.2**	**35.0**
with ChatGPT 3.5	12.8	23.8	16.6	16.7	29.2	28.3	32.5
CaseLink with Llama	**18.3**	**33.9**	**23.7**	**24.4**	**41.4**	**39.2**	**44.0***
with ChatGPT 3.5	17.1	31.8	22.3	22.7	34.1	33.3	38.6*
COLIEE 2024							
PromptCase with Llama	**20.2**	**25.8**	**22.6**	**24.7**	**35.0**	**33.8**	**39.0**
with ChatGPT 3.5	18.3	23.4	20.6	22.6	30.8	30.0	34.6
CaseGNN with Llama	**18.4**	**23.6**	**20.7**	**21.8**	**37.7**	**36.3**	**41.8***
with ChatGPT 3.5	16.4	21.0	18.4	19.4	35.1	33.6	38.2*
CaseLink with Llama	**23.3**	**29.8**	**26.2**	**28.5**	**44.9**	**43.5**	**49.7***
with ChatGPT 3.5	22.2	28.4	24.9	26.5	42.9	41.2	47.1*

Table 6: Results on the COLIEE 2022, 2023 and 2024 dataset with Llama plugged into the full pipeline. Significant differences on the NDCG@5 metric between the ChatGPT 3.5 and Llama setup within the same year using paired t-tests at $p < 0.05$ are marked with *. Bold indicates best result for this metric within the same dataset and approach.

data representations did not result in any improvements. Furthermore, we tested an open LLM alternative in the CaseLink pipeline as part of PromptCase, addressing the challenges we faced with the commercial OpenAI GPT model used in the original paper. Our findings suggest that Llama is a viable alternative, supporting the reproducibility of LLM-based methods in the field and even resulting in consistently better performance.

We share our repository that we optimized for reproducibility of the approach with the community and hope that the insights from our experiments will support future research in IR.

8 Limitations

While we expanded on the experimental setup to assess the reliability and generalizability of the original work, several limitations remain and we will briefly discuss them in this section.

First, our experiments focus exclusively on legal case retrieval, which is a specialized subfield within the broader domains of legal retrieval and professional search. As a result, the insights we gained from our study may not generalize to other areas of legal information retrieval.

Second, the three benchmark datasets used in our experiments consist solely of cases from the Canadian Federal Court. This limits the applicability of the findings to other legal systems, as it remains unclear how well the approach would transfer to datasets from different jurisdictions, for example from Germany or China.

Third, while we have taken an initial step towards more sophisticated graph modeling, several directions for further improvements remain. For example, edges could be directed based on BM25 scores, or these scores could be included as edge weights. That can be based on the heterogeneous graph representations introduced in our study as they make it easier to model this information.

Finally, our work does not account for the dynamic nature of legal data. The datasets and data modeling used in this study do not capture how legal cases and references evolve over time, which is an important aspect of real-world case law. Future research should explore approaches that consider temporal dynamics.

9 Ethical Considerations

We do not identify any immediate ethical concerns arising from our work. The datasets used in our study are made available for research purposes under the condition of signing a memorandum, which we did before starting the experiments. Similarly, our use of the Llama LLM is in line with the model's license. To make our work transparent, we make our implementations and experimental artifacts publicly available to the community, to support the reproducibility of our results.

From a broader perspective, we acknowledge that while legal retrieval tools can help in reducing the workload of legal professionals, they may also influence decision-making and critical legal reasoning. Over-reliance on AI-driven retrieval could shift responsibility away from human legal experts, which raises concerns about accountability in case selection and legal interpretation.

Acknowledgments

We would like to thank Yanran Tang for supporting our reproducibility study by promptly answering our questions.

References

[1] AI@Meta. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[2] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 67–93.

[3] Steven M Barkan, Roy M Mersky, and Donald J Dunn. 2009. *Fundamentals of legal research*. Foundation Press, 9th Edition.

[4] Keping Bi, Xiaojie Sun, Jiafeng Guo, and Xueqi Cheng. 2024. Reproducibility Analysis and Enhancements for Multi-aspect Dense Retriever with Aspect Learning. In *European Conference on Information Retrieval*. Springer, 194–209.

[5] 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[6] Gregor Donabauer and Udo Kruschwitz. 2023. Exploring Fake News Detection with Heterogeneous Social Media Context Graphs. In *European Conference on Information Retrieval*. Springer, 396–405.

[7] Gregor Donabauer, Anca Rath, Aila Caplunik-Pratsch, Anja Eichner, Jürgen Fritsch, Martin Kieninger, Susanne Gaube, Wulf Schneider-Brachert, Udo Kruschwitz, and Bärbel Kieninger. 2025. AI modeling for outbreak prediction: A graph-neural-network approach for identifying vancomycin-resistant enterococcus carriers. *PLOS Digital Health* 4, 4 (2025), e0000821.

[8] 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)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6472–6485. doi:10.18653/v1/2024.acl-long.350

[9] Tobias Fink. 2022. Graph-Enhanced Document Representation for Court Case Retrieval. In *European Conference on Information Retrieval*. Springer, 480–487.

[10] 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* (Braga, Portugal) *(ICAIL '23)*. Association for Computing Machinery, New York, NY, USA, 472–480. doi:10.1145/3594536.3595176

[11] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024. In *New Frontiers in Artificial Intelligence*, Toyotaro Suzumura and Mayumi Bono (Eds.). Springer Nature Singapore, Singapore, 109–124.

[12] Artur Guimarães, João Magalhães, and Bruno Martins. 2025. A Reproducibility Study on Consistent LLM Reasoning for Natural Language Inference over Clinical Trials. In *European Conference on Information Retrieval*. Springer, 48–63.

[13] Hanjo Hamann. 2019. The German federal courts dataset 1950–2019: From paper archives to linked open data. *Journal of empirical legal studies* 16, 3 (2019), 671–688.

[14] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2704–2710. doi:10.1145/3366423.3380027

[15] 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*, Yasufumi Takama, Katsutoshi Yada, Ken Satoh, and Sachiyo Arai (Eds.). Springer Nature Switzerland, Cham, 51–67.

[16] Osman Alperen Koraş, Jörg Schlötterer, and Christin Seifert. 2024. A Second Look on BASS–Boosting Abstractive Summarization with Unified Semantic Graphs: A Replication Study. In *European Conference on Information Retrieval*. Springer, 99–114.

[17] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. 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* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1035–1044. doi:10.1145/3539618.3591761

[18] Yixiao Ma, Yueyue Wu, Qingyao Ai, Yiqun Liu, Yunqiu Shao, Min Zhang, and Shaoping Ma. 2023. Incorporating Structural Information into Legal Case Retrieval. *ACM Trans. Inf. Syst.* 42, 2, Article 40 (Nov. 2023), 28 pages. doi:10.1145/3609796

[19] David Rau and Jaap Kamps. 2024. Query Generation Using Large Language Models: A Reproducibility Study of Unsupervised Passage Reranking. In *European Conference on Information Retrieval*. Springer, 226–239.

[20] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[21] 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 *IJCAI*. 3501–3507.

[22] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2023. Understanding Relevance Judgments in Legal Case Retrieval. *ACM Transactions on Information Systems* 41, 3 (2023), 1–32.

[23] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2021. Investigating user behavior in legal case retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 962–972.

[24] Faisal Shehzad and Dietmar Jannach. 2024. Performance Comparison of Session-Based Recommendation Algorithms Based on GNNs. In *European Conference on Information Retrieval*. Springer, 115–131.

[25] Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Negreanu, and Gust Verbruggen. 2023. CodeFusion: A Pre-trained Diffusion Model for Code Generation. arXiv:2310.17680 [cs.SE] https://arxiv.org/abs/2310.17680

[26] Aman Sinha, Priyanshu Raj Mall, and Dwaipayan Roy. 2024. Exploring the Nexus Between Retrievability and Query Generation Strategies. In *European Conference on Information Retrieval*. Springer, 177–193.

[27] Moritz Staudinger, Wojciech Kusa, Florina Piroi, Aldo Lipani, and Allan Hanbury. 2024. A Reproducibility and Generalizability Study of Large Language Models for Query Generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 186–196.

[28] Yanran Tang, Ruihong Qiu, and Xue Li. 2024. Prompt-Based Effective Input Reformulation for Legal Case Retrieval. In *Databases Theory and Applications*, Zhifeng Bao, Renata Borovica-Gajic, Ruihong Qiu, Farhana Choudhury, and Zhengyi Yang (Eds.). Springer Nature Switzerland, Cham, 87–100.

[29] Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024. CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs. In *Advances in Information Retrieval*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer Nature Switzerland, Cham, 80–95.

[30] Yanran Tang, Ruihong Qiu, Hongzhi Yin, Xue Li, and Zi Huang. 2024. CaseLink: Inductive Graph Learning for Legal Case Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2199–2209. doi:10.1145/3626772.3657693

[31] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[32] Suzan Verberne. 2024. Professional Search. In *Information Retrieval: Advanced Topics and Techniques*, Omar Alonso and Ricardo Baeza-Yates (Eds.). Association for Computing Machinery, New York, NY, USA, 501–514. https://doi.org/10.1145/3674127.3674141

[33] Suzan Verberne, Evangelos Kanoulas, Gineke Wiggers, Florina Piroi, and Arjen P de Vries. 2023. ECIR 2023 Workshop: Legal Information Retrieval. In *European Conference on Information Retrieval*. Springer, 412–419.

[34] Shuai Wang and Guido Zuccon. 2023. Balanced Topic Aware Sampling for Effective Dense Retriever: A Reproducibility Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2542–2551. doi:10.1145/3539618.3591915