Information Processing and Management 59 (2022) 103051

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Query Generation and Buffer Mechanism: Towards a better conversational agent for legal case retrieval☆

Bulou Liu [a], Yueyue Wu [a], Fan Zhang [b], Yiqun Liu [a,*], Zhihong Wang [a], Chenliang Li [c], Min Zhang [a], Shaoping Ma [a]

[a] Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China
[b] School of Information Management, Wuhan University, Bayi Rd., Wuhan, 430075, Hubei, China
[c] School of Cyber Science and Engineering, Wuhan University, Bayi Rd., Wuhan, 430075, Hubei, China

ARTICLE INFO

Keywords:
Legal case retrieval
Conversational agent
Workflow
User study

ABSTRACT

In legal case retrieval, existing work has shown that human-mediated conversational search can improve users' search experience. In practice, a suitable workflow can provide guidelines for constructing a machine-mediated agent replacing of human agents. Therefore, we conduct a comparison analysis and summarize two challenges when directly applying the conversational agent workflow in web search to legal case retrieval: (1) It is complex for agents to express their understanding of users' information need. (2) Selecting a candidate case from the SERPs is more difficult for agents, especially at the early stage of the search process. To tackle these challenges, we propose a suitable conversational agent workflow in legal case retrieval, which contains two additional key modules compared with that in web search: **Query Generation** and **Buffer Mechanism**. A controlled user experiment with three control groups, using the whole workflow or removing one of these two modules, is conducted. The results demonstrate that the proposed workflow can actually support conversational agents working more efficiently, and help users save search effort, leading to higher search success and satisfaction for legal case retrieval. We further construct a large-scale dataset and provide guidance on the machine-mediated conversational search system for legal case retrieval.

1. Introduction

In recent years, legal case retrieval has attracted much attention in both legal information processing and the IR research community. It aims to retrieve supporting cases for a given query case and constitutes an essential component of a legal information system. In practice, prior cases decided in courts of law are primary legal materials in various law systems, in addition to status. For instance, prior cases are fundamental for a lawyer when preparing the legal reasoning in the countries that follow the common law system. In the countries following the civil law system, constructing legal search systems is also increasingly important because it promotes the consistency in application of law and the supervision on judges (Hamann, 2019). Existing works show that an automatic system not only performs the search tasks with higher performance than lawyers, but also completes them more efficiently (McGinnis

☆ This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 62002194), Beijing Academy of Artificial Intelligence (BAAI), and Tsinghua University Guoqiang Research Institute.

* Corresponding author.
E-mail addresses: lbl20@mails.tsinghua.edu.cn (B. Liu), yiqunliu@tsinghua.edu.cn (Y. Liu).
URL: https://www.thuir.cn/group/~YQLiu/ (Y. Liu).

https://doi.org/10.1016/j.ipm.2022.103051
Received 1 March 2022; Received in revised form 22 July 2022; Accepted 25 July 2022

& Wasick, 2014). Using existing legal case retrieval system, users need to issue queries to express their complex information needs (Ferrer, Hernández, & Boulat, 2014; McGinnis & Pearce, 2019), which is a difficult task for them (Shao et al., 2021). It is much more prominent in legal case retrieval comparing to ordinary web search that information need is more complex and legal domain knowledge is required during search (Shao et al., 2021).

Conversational search is an emerging paradigm in IR, referring to the phenomenon that an information retrieval system proactively refines users' requests and search results iteratively and interactively (Radlinski & Craswell, 2017). Such a system can help users express their information needs better (Radlinski & Craswell, 2017) and improve search performance (Carterette, Kanoulas, Hall, & Clough, 2014), especially for complex and exploratory search tasks (Belkin, Cool, Stein, & Thiel, 1995; Solomon, 1997). Conversational search paradigm has been adopted in various search scenarios, such as web search (Zamani & Craswell, 2020), product search (Zhang, Chen, Ai, Yang, & Croft, 2018), academic search (Balog et al., 2020) and so on. In legal case retrieval, a typical complex information search scenario, studies have concluded that conversational search paradigm can help improve users' search experience, in terms of query formulation, result examination, users' satisfaction and search success (Liu et al., 2021).

The most popular way to implement a conversational search system is to add an intermediate agent between the traditional search engine (system) and the user (Ren et al., 2021; Thomas, McDuff, Czerwinski, & Craswell, 2017), namely the wizard-of-oz approach. The agent aims to understand the search intent expressed by the user in natural language and submits queries to the system. Then it collects candidates from the system and reorganizes them as answers for users. Due to the fact that there exists no available conversational legal case retrieval system, Liu et al. (2021) recruited legal experts as intermediary agents in the wizard-of-oz approach, and verified the effectiveness of conversational agents. And a suitable workflow can provide guidelines for constructing machine-mediated agents replacing of human agents. Therefore, it is important to construct a suitable workflow for the agent in conversational legal case retrieval. In general web search scenarios, existing studies have conceptualized the conversational search process and proposed frameworks for the development of practical systems (Azzopardi, Dubiel, Halvey, & Dalton, 2018; Radlinski & Craswell, 2017; Ren et al., 2021). Specifically, Ren et al. (2021) suggested a conversational agent workflow and provided a guideline on how to build datasets and implement computation models. However, legal case retrieval differs from general web search in various aspects, such as the needs for data authority (Arewa, 2006), the definition of relevance (Ferrer et al., 2014; Fleiss, 1971), the user behavior (Shao et al., 2021) and so on. The differences suggest that we should not directly adopt the workflow from general web search to legal case retrieval, which leads to two research questions:

- **RQ1:** Where do the differences lie in search process between legal case retrieval and general web search in conversational search paradigm?
- **RQ2:** How can we construct a suitable conversational agent workflow for legal case retrieval?

To address the first question, we first investigate the differences of search process, relating to user behavior and agent action, between legal case retrieval and general web search in conversational search paradigm. Specifically, we use two public conversational search behavior datasets (i.e., Ren et al., 2021 for general web search and Liu et al., 2021 for legal case retrieval) for the comparative analysis. Based on the differences summarized, we find that two serious challenges are confronted when applying existing conversational agent workflow in general web search to legal case retrieval directly:

- **Challenge 1.** It is more complex for the agent to express his understanding of users' information need by organizing queries.
- **Challenge 2.** It is more difficult for agents to distinguish whether a candidate case is relevant, especially at the early stages of the retrieval process.

To tackle these two challenges and address the second research question, we propose a suitable conversational agent workflow for legal case retrieval, which contains two additional key modules compared with the existing workflow for general web search: **Query Generation** and **Buffer Mechanism**. Here, Query Generation denotes that while constructing and submitting queries to the system, agents can not only extract keywords from the conversation history, but also generate new phrases which are not included in the conversation but can express agents' understanding of the information need appropriately. It would be beneficial especially when it is difficult for users to use accurate legalese. As to Buffer Mechanism, we set up a buffer for agents to save all the cases they have read and return cases to users from the search engine result pages (SERPs) of all the queries (related to a variety of possible charges) in the conversation. We then investigate whether these two modules can help the conversational agents work better, which leads to the second research question:

- **RQ3:** How Query Generation and Buffer Mechanism in the proposed workflow affect agent actions?

In additional to investigate the changes of agent actions, we can further investigate whether these two modules can improve users' legal case retrieval experience, raising the third research question:

- **RQ4:** How Query Generation and Buffer Mechanism in the proposed workflow affect user search behavior and outcome?

To shed light on the third and fourth research questions, we design a controlled lab-based user study (including 157 tasks) with three different experimental settings to examine the effectiveness of the two modules. Specifically, the agents followed the workflow proposed in Section 4 in the first group (denoted as "Whole") which is the baseline group for comparison. Then we restrict the conversational agent action in two ways. In the second group (denoted as "-QG"), we remove Query Generation from our proposed workflow and the agents cannot generate any phrase by themselves. In the third group (denoted as "-BM"), we exclude Buffer Mechanism from our proposed workflow and the agents cannot select cases from the SERPs of the previous queries. Except

for the restrictions on agent actions, these two groups of user study follow the setting of the whole workflow (i.e., the "Whole" group) for the fair comparison. We logged rich interaction behavioral data in the search process including conversation contents, agents' queries, clicks, etc. We also collected users' and agents' self-reported feedback and relevance assessments in the study.

With the collected data, we systematically analyze the effects of the two modules. We find that agents ask less clarifying questions and achieve higher query performance with Query Generation. Therefore, users can save efforts in answering clarifying questions. Buffer Mechanism save agents' efforts in examining candidate cases, so that users can browse fewer candidate cases. In general, both of the modules can help users achieve higher satisfaction. That is to say, our proposed workflow can help the conversational agent work better and improve users' search experience in legal case retrieval.

We provide further practical implications and discussions on the machine-mediated conversational search system for legal case retrieval. Specifically, in our proposed workflow, there are five kinds of agent actions, namely conversation examining, intent understanding, query extraction and generation, buffer filling and buffer examining. Therefore, we formalize five computational tasks corresponding to the five kinds of actions to construct a machine agent for conversational legal case retrieval. We then construct a large-scale dataset following the workflow and provide guidance for developing conversational legal case retrieval systems.

In summary, the main contributions of this article are as follows:

- We investigate the differences of search process between legal case retrieval and general web search in conversational search paradigm and find that two serious challenges are confronted when applying existing conversational agent workflow in general web search to legal case retrieval directly.
- We propose a suitable conversational agent workflow for legal case retrieval, which contains two additional key modules: Query Generation and Buffer Mechanism and show that both of the modules can help users achieve higher satisfaction by a controlled lab-based user study.
- We construct a large-scale dataset for model training and evaluation following the workflow and provide further guidelines for constructing a machine-mediated agent replacing of human agent.

2. Related work

2.1. Legal case retrieval

Legal case retrieval is a specialized IR task, which is different from general web search in various aspects. Ferrer et al. (2014) stated that relevance in legal domain implies applicable legal criteria rather than a simple frequency calculation. Van Opijnen and Santos (2017) pointed out several shortcomings of general IR in the legal domain and developed a unique framework with six 'dimensions' for the concept of relevance in legal information retrieval. Turtle (1995) discussed some distinctive aspects of legal retrieval and concluded that the characteristics of legal documents that are different from materials of other domains, including professional legal expressions, the special logical structures and so on. Moreover, legal information services (e.g., Westlaw) meet the users' demand considering data authority and legal effect, which are connected to the nature of legal practice (Doyle, 1992). Additionally, most retrieved results in legal case retrieval are semi-structured case documents instead of unstructured web pages. Shao et al. (2021) investigated user behavior in legal case retrieval. They found that compared with general web search, users of legal case retrieval devote more search effort, incorporate domain-specific knowledge, and appear to be more patient and cautious.

Several approaches have been explored in previous research of legal IR, including knowledge engineering-based techniques and NLP-based methods. For instance, Rose and Belew (1989) combined symbolic and connectionist artificial intelligence techniques to integrate both symbolic and sub-symbolic information in legal domain. Saravanan, Ravindran, and Raman (2009) developed a legal knowledge-based framework to overcome synonymy and ambivalence of words in query process and enhance the user's query for retrieving truly relevant legal judgments. Jackson, Al-Kofahi, Tyrrell, and Vachher (2003) employed a combination of information retrieval and machine learning techniques to link new cases to related documents that may be impacted by new cases. They presented a complete system that combines partial parsing techniques with domain knowledge and discourse analysis to extract information from the free text of court opinions. However, these existing legal case retrieval systems still followed a traditional search paradigm in which users issued keyword-based queries to describe their information needs (Ferrer et al., 2014; McGinnis & Pearce, 2019). Shao, Mao et al. (2020) proposed a BERT-based neural network to model paragraph-level interactions for legal case retrieval. And Shao, Mao et al. (2020) and Ma et al. (2021) suggested that BERT-based neural networks improved the performance of the legal case retrieval task significantly.

2.2. Conversational search

Several user studies have been conducted to explore the value of conversational search system. Vtyurina, Savenkov, Agichtein, and Clarke (2017) recruited 21 participants and asked each of them to complete 3 complex information tasks conversing with three different conversational systems: an existing commercial intelligent assistant, a human expert and a perceived automatic system (a human disguised as an automatic system). They pointed out that users do not have biases against perceived automatic systems. And users are glad to use them as long as users' requirement of accuracy satisfied. Additionally, human and perceived automatic systems could understand users' partially stated questions better. Trippas, Spina, Cavedon, Joho, and Sanderson (2018) conducted a laboratory-based observational study where users completed search tasks with verbal communication. Their results

Table 1
Properties of the conversational search datasets of general web search and legal case retrieval.

Dataset	User domain	Language	#Task	Log detail
WISE (Ren et al., 2021)	Non-specific	Chinese	705	conversation, response label, query, candidate query, candidate document, selected query, selected document
CLCR (Liu et al., 2021)	Legal	Chinese	55	conversation, query, candidate document, clicked document, selected document

highlighted that conversational search has increased complexity and interactivity compared with traditional IR system. Referring to the aforementioned research, conversational search system provides opportunities to improve search quality, especially for complex and exploratory search tasks.

There are several frameworks aimed to address a specific aspect of conversational search. Zamani, Dumais, Craswell, Bennett, and Lueck (2020) focused on generating clarifying questions for open-domain search tasks to address ambiguous search queries. Nguyen et al. (2016) developed a new, large-scale dataset with questions from real user queries and human generated answers to inspire work in reading comprehension and question answering. There are also attempts to implement a complete model or framework of a conversational search system. Radlinski and Craswell (2017) suggested five desirable properties of a conversational information retrieval system, with which system can allow users to answer information needs in a natural and efficient manner. They also presented a theoretical model that satisfies these desirable properties and provided the framework for a conversational search system. Azzopardi et al. (2018) developed a conceptual framework of actions and intents of users and systems to explain their roles when users explore the search space and resolve their information need. They suggested a conceptualization of the conversational search process which provides a framework for development of conversational search agents. Wang and Ai (2021) proposed a risk-aware conversational search agent model to balance the risk of answering user's query and asking clarifying questions. Hashemi, Zamani, and Croft (2020) enriched the representations learned by Transformer networks by using a novel attention mechanism from external information sources that weights each term in the conversation. Yu, Liu, Xiong, Feng, and Liu (2021) presented a Conversational Dense Retrieval system, ConvDR, that learns contextualized embeddings for multi-turn conversational queries and retrieves documents solely using embedding dot products. Ren et al. (2021) proposed a workflow and a modular end-to-end neural architecture to construct conversational agents for general web search, which models it as six sub-tasks to improve the performance on the response generation task. Their work provides a conceptualization of the conversational search process and a framework for the development of practical conversational systems for general web search. Compared to these studies, our work is the first to investigate how to construct a suitable conversational agent workflow for legal case retrieval.

Liu et al. (2021) conducted a laboratory-based user study to investigate whether conversational search paradigm can be adopted to improve users' legal case retrieval experience. They found that users achieve higher satisfaction and success in conversational legal case retrieval, especially when they lack sufficient domain knowledge. Compared to this research, the improvements of our work are as follows: (1) We further investigate the differences of search process between legal case retrieval and general web search in conversational search paradigm. (2) We construct a suitable conversational agent workflow for legal case retrieval and show the effectiveness of the key modules in our workflow. (3) Based on our workflow, we provide guidance for model designing and construct a large-scale conversational legal case retrieval dataset for model training. In summary, the previous work has demonstrated the significance of building a conversational legal case retrieval system and our work investigated how to construct conversational legal case retrieval systems.

3. Preliminary study

To address **RQ1**, we first conduct a preliminary study to investigate the differences of search process between legal case retrieval and general web search in conversational search paradigm. Following these comparisons, we then summarize two unique challenges when constructing a conversational agent for legal case retrieval.

3.1. Datasets for comparison

We utilized two public conversational search datasets with search behaviors to examine the differences in terms of user behaviors and agent actions between legal case retrieval and general web search. As to the general web search, we choose the conversational search dataset (namely WISE) from Ren et al. (2021). For the legal case retrieval, the dataset generated from Liu et al. (2021) is used instead (namely CLCR). Table 1 shows the properties of these datasets. The two datasets are both in Chinese and built in a wizard-of-oz fashion. Both datasets include rich behavioral features. In this sense, we can perform a comprehensive comparison towards the search process between conversational general web search and conversational legal case retrieval. Note that CLCR does not contain agent response labels of each utterance (Ren et al., 2021) (e.g., clarify: helping the users to clarify their intents whenever unclear, answer: providing retrieved cases). Hence, we asked one of the authors to manually label all the utterances in CLCR according to the label taxonomy in WISE.

Table 2
Agent-user interaction behavioral measures in legal/web datasets. "*/**/***" indicates the difference in the measure is statistically significant at $p < 0.05/0.01/0.001$ level.

Group	Measure	WISE	CLCR	sig
User	#Utterances	10.60	5.455	***
	#Avg words per utterance	11.27	23.37	***
	#Avg words per conversation	119.4	127.5	–
Agent-Overall	#Utterances	8.651	5.509	***
	%clarify	10.08	73.42	***
	%answer	66.52	25.92	***
	%request-rephrase	0.920	0.670	–
	%recommend	0.340	0	–
	%chitchat	22.13	0	***
Agent-Clarify	#Questions	0.868	4.000	***
	#Avg question words per conversation	17.37	68.64	***
	#Avg words per question	20.01	17.16	***
	%clarify-yes-no in clarify	29.58	35.29	–
	%clarify-choice in clarify	51.96	4.980	***
	%clarify-open in clarify	18.46	59.73	***
Agent-Answer	#Answers	5.730	1.345	***
	%answer-link in answer	7.892	97.30	***

3.2. Comparison of behavioral measures

As we mentioned in the introduction, the agent understands the search intent expressed by the user in natural language and submits queries to the system. Then it collects candidate results from the system and organizes them into answers for users. It mainly contains two types of actions: interactions with users and interactions with the system. So we investigate the differences of this search process from two aspects: Agent-User interaction and Agent-System interaction. Mann–Whitney U test (Mann & Whitney, 1947) instead of the t-test is conducted since most of the variables have a non-normal distribution.

Agent-user interaction. In this aspect, we first consider two types of behavior measures: user behaviors (User), agent actions (Agent-Overall). In addition, when the agent interacts with the user, the agent needs to understand the user's search intent and organize the relevant results. Consequently, we further categorize the agent actions into two fine-grained groups: the actions related to agent's clarifying question (Agent-Clarify) and the actions related to the agent's answer (Agent-Answer). The results are shown in Table 2.

User Behaviors. Regarding user behaviors, we focus on three conversation-related indicators: the average number of utterances that are issued by a user in a conversation session (#Utterances), the average number of words that are generated by a user in each of her utterances (#Avg words per utterance) and the average number of words that are generated by a user during a conversation session (#Avg words per conversation). Specifically, as to the total word number of a conversation, we find that there is no significant difference across the two scenarios. However, a user issues on average 5.455 utterances in conversational legal case retrieval, which is significant less than that in conversational general web search (5.455 vs 10.60). That is, a user issues longer utterances each time in conversational legal case retrieval than that in conversational general web search (11.27 vs 23.37). In other words, a user needs to include more information to express the information need. This indicates that the expression of information need in conversational legal case retrieval is more complex and difficult.

Agent-Overall. As for agent actions, we observe that an agent also takes less utterances to finish a user search task than that in conversational general web search (5.509 vs 8.651). More specifically, we focus on the proportion of various types of responses. On the one hand, there are less types of response from conversational agents in legal case retrieval than those in general web search. Without "chitchat-type" and "recommend-type" response, a legal agent mainly concentrates on satisfying users' specific information needs rather than making users staying longer. On the other hand, the proportion of "clarify-type" responses becomes significantly larger in conversational legal case retrieval (from 10.08% to 73.42%). In contrast, the proportion of "answer-type" responses significantly decreases in conversational legal case retrieval (from 66.52% to 25.92%), which also indicates that the information needs are more complex in conversational legal case retrieval. Therefore, a legal agent needs to ask more clarifying questions to understand the user's search intent before replying an answer. Based on the next two groups of behavioral measures (i.e., Agent-Clarify and Agent-Answer), we investigate the difference of agent-user interactions in the two search scenarios more deeply.

Agent-Clarify. In conversational legal case retrieval, an agent asks 4 clarifying questions in one search session on average, which is significant more than those in general web search (4 vs 0.868). The number of total word usage in clarifying questions of a search task in conversational legal case retrieval is also more than those in conversational general web search (68.64 vs 17.37). What is more, we find that one legal clarifying question contains 17.16 words on average, which are significantly less than that in general web search (20.01 words). So we further compare the proportion of three types of clarifying questions between the two scenarios. Firstly, we observe that a legal agent is more inclined to ask "open-type" clarifying questions in legal case retrieval (from 18.46% to 59.73%). And the proportion of "choice-type" clarifying questions decreases significantly in conversational legal case retrieval. These explain why legal clarifying questions are shorter and indicate that it is harder for agents to narrow down the users' search intent in conversational legal case retrieval than that in conversational general web search.

Table 3
Agent-system interaction behavioral measures in legal/web datasets. "*/**/***" indicates the difference in the measure is statistically significant at $p < 0.05/0.01/0.001$ level.

Group	Measure	WISE	CLCR	sig
Query formulation	#Queries	8.651	2.691	***
	#Avg words per query	1.930	3.081	***
	%generalization	22.61	5.376	***
	%specification	16.70	17.20	–
	%substitution	60.70	77.42	**
	%term-level generation	0	39.73	***
	%query-level generation	0	65.19	***
Candidate selection	%candidate->answer	14.74	12.36	*
	%selected candidates not from last turn	0	7.382	***

Agent-Answer. Regarding the actions related to the agent's answer, we firstly find that agents reply less number of answers in conversational legal case retrieval. This is because users' information needs of legal case retrieval are more focused than those in general web search. In conversational general web search, users usually would like to learn about one entity from multiple aspects or other related entities in one session. However, in conversational legal case retrieval, users only concentrate on one specific situation in a session. Furthermore, different from conversational general web search, almost all answers of conversational legal case retrieval are links of cases. Therefore, a legal agent only needs to choose the answer cases instead of organizing answer passages or words by themselves.

Agent-system interaction. When the agent interacts with the system, the agent first submits queries to the system, and then collects candidate results from the SERPs. Therefore, we also divide the agent action measures into two fine-grained groups: query formulation and candidate selection. The results are shown in Table 3.

Query Formulation. First, conversational agents issue less queries to the system in legal case retrieval than those in general web search significantly (2.691 vs 8.651). This is also because the information need of legal case retrieval is more focused than those in general web search. Moreover, a legal agent appears to formulate longer queries in conversational legal case retrieval. This indicates that it is complex and difficult not only for users to express their information need, but also for agents to express their understanding of users' information need by organizing queries in conversational legal case retrieval. Second, we investigate the query reformulation type which is automatically determined following the previous work (Mao, Liu, Kando, Zhang, & Ma, 2018). We observe that compared with general web search, the type *substitution* comprises a larger proportion (77.42% vs 60.70%), indicating that a legal agent needs to make some trials to find an appropriate query. Third, we investigate whether there are some query words that are not mentioned in the preceding conversation. To be specific, *%term-level generation* denotes the proportion of query words not mentioned in conversation history and *%query-level generation* denotes the proportion of queries which contain at least one word not mentioned in conversation history. We find that all the query words are included in the web search conversations according to the corresponding workflow setting of the latter. However, in conversation legal case retrieval, a legal agent generates several query words (39.73%) by themselves and more than half of the queries contain at least one word not mentioned in conversation history. This indicates that sometimes a user is unable to express her information need in terms of accurate legalese so that agents need to formulate some queries by themselves in legal case retrieval.

Candidate Selection. Regarding candidate selection, we investigate how the agents select candidates from the system to generate answers. It is worth noting that in general web search, candidates include suggested queries and retrieved documents in SERPs. On the contrary, the retrieved candidates are legal cases for legal case retrieval. On the one hand, we find that 12.36% of retrieved candidates in legal case retrieval are identified by the agent as answers (*%candidate->answer*), which is significantly less than the proportion of that in general web search (14.74%). This illustrates that agents need to spend more efforts in examining the candidates to generate an answer for a user. On the other hand, we check whether there are some candidates which are selected to compose answers but not from the SERP of the last query (*%selected candidates not from last turn*). It is obvious that all the selected candidates are from the last SERP in general web search. However, 7.382% of selected cases are not from the last turn of query in conversational legal case retrieval. This indicates that it is more difficult for agents to distinguish whether a candidate case should be chosen exactly in legal case retrieval. A legal agent sometimes needs to issue more queries and compare the candidates of the current turn with the candidates from the previous turns.

3.3. Summary

Regarding **RQ1**, we summarize the difference between conversational legal case retrieval and conversational general web search in terms of agent-user interaction process and agent-system interaction process. As for agent-user interactions, we find that agent response types of legal case retrieval are less than that of general web search (without chitchat, recommend, etc.). And a legal agent needs to ask more clarifying questions (especially "open-type" questions) to understand a user's search intent. What is more, almost all answers of conversational legal case retrieval are links the of cases according to the task setting. As for agent-system interactions, compared with general web search, a legal agent needs to issue longer queries and generate some query words by themselves in legal case retrieval. In additional, a legal agent needs to examine more candidates to generate an answer and sometimes cannot distinguish whether a candidate case should be chosen exactly without comparison. Therefore, we summarize two challenges that we face when constructing a conversational agent for legal case retrieval:

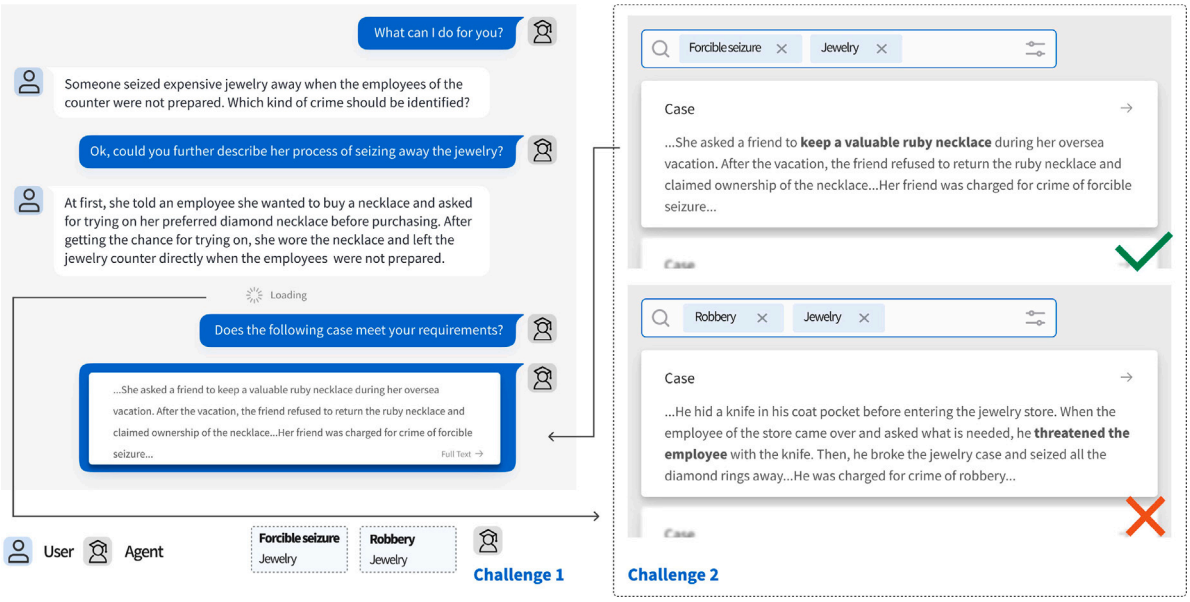

Fig. 1. An example about the two challenges in conversational legal case retrieval. On the one hand, the agents need to utilize accurate legalese like forcible seizure and robbery (corresponding the **Challenge 1**). On the other hand, the agents need to compare cases from different queries related to a variety of possible charges (corresponding the **Challenge 2**).

- **Challenge 1.** It is more complex for the agent to express his understanding of users' information need by organizing queries.
- **Challenge 2.** It is more difficult for agents to distinguish whether a candidate case should be chosen exactly.

Fig. 1 shows an example about these two challenges in conversational legal case retrieval. Based on the conversation, on the one hand, the agents need to utilize accurate legalese (e.g., forcible seizure and robbery) and some facts to formulate the queries. So it is more complex for the agent to express his understanding of users' information. On the other hand, to make relevance judgement on a case, the agents usually need to check about cases not only about one charge, but also cover a variety of possible charges related to the current fact description. So it is more difficult for agents to distinguish whether a candidate case should be chosen exactly.

4. Workflow

To tackle the challenges mentioned in Section 3 and address **RQ2**, we propose a suitable conversational agent workflow for legal case retrieval, which is shown in Fig. 2. Specifically, we first summarize the differences with general web search while constructing a conversational agent for legal case retrieval, including the changes of output form and two additional modules: **Query Generation** and **Buffer Mechanism**. Then we describe the detailed process of the proposed workflow.

4.1. Differences with general web search

Regarding **RQ2**, there are four differences to construct a conversational agent for legal case retrieval against the general web search based on the findings from Section 3. We highlight and label them correspondingly in Fig. 2. According to the setting of legal case retrieval, a legal agent only needs to focus on understanding users' specific information need and retrieving relevant cases for them. Therefore, the first two differences are about the changes of output form:

- **Difference 1.** There are only two conversational agents' response types: "clarify" and "answer" in legal case retrieval, without "chitchat", "recommend" and so on.
- **Difference 2.** The answers of legal case retrieval only contain legal case documents, without processed passages, entities and other types of information.

The other two differences are two additional modules devised to tackle the two challenges summarized in Section 3. Regarding **Challenge 1**, one main problem is that it is difficult for users to express their information needs by accurate legalese. Therefore, a legal agent need to recruit some new query words based on her expertise, raising the third difference as follows:

- **Difference 3.** Conversational agents for legal case retrieval need one additional module: Query Generation. It denotes that while constructing queries and submitting them to the system, the agents can not only extract some keywords from the conversation history, but also propose some phrases which are not mentioned in the conversation but can express agents' understanding of the information need appropriately.

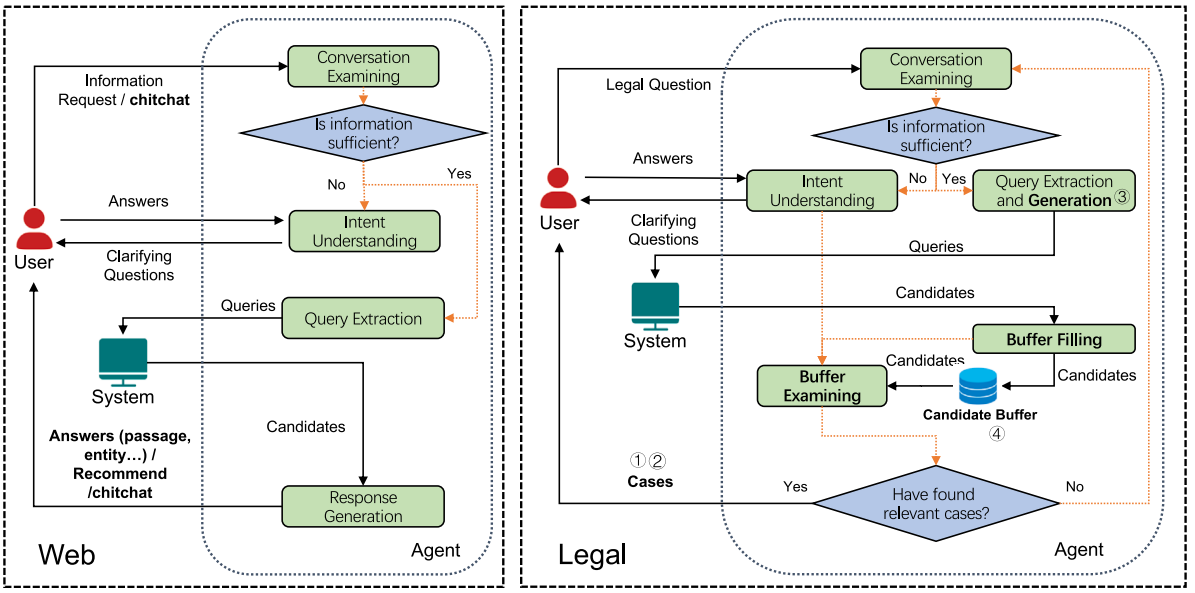

Fig. 2. Comparison of conversational agents' workflow for general web search (left) and our proposed workflow for legal case retrieval (right). We highlight and label the four differences between them corresponding to Section 4.1. The dotted lines represent the action flow and the solid lines represent the data flow.

Specifically, with Query Generation, the agents can firstly utilize accurate legalese (such as forcible seizure and robbery in Fig. 1) to more precisely limit the scope of cases returned by the system. This can help agents focus on the cases just related to specific and possibly related charges. And then the agents can extract facts (such as jewelry in Fig. 1) from the conversations to find similar cases more accurately.

As for **Challenge 2**, it is more difficult for agents to distinguish whether a candidate case is relevant to the query case exactly in legal case retrieval. A legal agent sometimes needs to issue more queries to collect more candidates for comparison, raising the fourth difference as follows:

- **Difference 4.** Conversational agents for legal case retrieval need another additional module: Buffer Mechanism. It denotes that an agent has a buffer to save all the cases it has read and selects cases to users from the search engine result pages (SERPs) of all the queries (related to a variety of possible charges) in the conversation.

Specifically, with Buffer Mechanism, the agents can submit a group of queries to cover a variety of possible charges related to the current fact description. And then the agents can compare all the possibly relevant cases from different queries with facts in the conversation and make more convincing relevance judgement.

4.2. Whole process

Based on the differences summarized in Section 4.1, we introduce the detailed process of the agent workflow as follows:

1. **Conversational Examining.** A legal agent examines the conversation history to determine whether the background information of the search task is sufficient.
2. **Intent Understanding.** If the legal agent does not acquire sufficient background information, a clarifying question is asked to understand the user's search intent better.
3. **Query Extraction and Generation.** If the legal agent thinks the background information is sufficient, a query is constructed, by not only extracting keyphrases from the conversation history but also generating phrases by themselves, and submitted to the system.
4. **Buffer Filling.** Following step 3, the agent examines the SERP according to the query and stores the candidates that may be relevant in the buffer.
5. **Buffer Examining.** After step 2 or 4, the agent examines the buffer to judge whether there are relevant cases to the query case. If relevant cases exist, the agent returns them to users as answers. If not, the agent goes back to execute step 1 again.

Following the workflow, an agent can repeat the above steps until a user finds enough information or her patience is exhausted.

Table 4
An example of two criminal tasks in different topics.

ID	Task 1	Task 2
Topic	Forcible rape	Fraud
Fact description	A was originally a male, but changed into a female after a sex change operation. Her physiology showed female characteristics, and her household registration and ID card information were subsequently changed to female. B forcibly had sex with A against A's will.	A lent a sum of money to B, deducted the deposit, interest, and service fee in advance from the principal amount lent, and required B to return the three aforementioned fees to another person's bank account designated by A, making the principal lending record consistent with the amount returned.
Legal issue	The subjects of the crime of rape.	Whether trap loans is constituted fraud?

5. User study

To investigate whether the two additional modules (Query Generation and Buffer Mechanism) can tackle the challenges summarized in Section 3, we conducted three groups of lab-based user study with 157 tasks. Specifically, the agents followed the workflow proposed in Section 4 in the first group (denoted as "Whole") which is the baseline group for comparison. Then we restrict the conversational agent action in two ways. In the second group (denoted as "-QG"), the conversational agents can only extract keyphrases from the conversational history to construct queries. Namely, we remove Query Generation from our proposed workflow and the agents cannot generate any phrase by themselves. In the third group (denoted as "-BM"), the conversational agents can only select cases from the SERP of the last query as the answer. In other words, we exclude Buffer Mechanism from our proposed workflow and the agents cannot select cases from the SERPs of the previous queries. Except for the restrictions on agent actions, these two groups of user study follow the setting of the whole workflow (i.e., the "Whole" group) for the fair comparison. In this section, we describe the details of the user study and the dataset we collected.

5.1. Tasks and participants

We collected 157 search tasks from legal practitioners' real information need via online forums and social networks, covering 3 legal domains: 53 civil tasks (involving "Inheritance", "Personality rights", "Contracts" and "Marriage" topics), 51 criminal tasks (involving "Robbery", "Fraud", "Bribery", "Forcible Rape" and "Traffic accident" topics) and 53 commercial tasks (involving "Company", "Expertise Bankruptcy" and "Insurance" topics). Each task contained a query case description and a legal issue. Users were expected to retrieve legal cases which may help to answer the issue question.

There were two kinds of participants: users and agents. As for users, we recruited 157 participants (41 males and 116 females) via online forums and social networks. They were all native Chinese speakers and college law students. All users had no previous experience with conversational search systems. Each user conducted three different cases with three different settings (i.e., "Whole", "-QG" and "-BM"), respectively. This ensures the set of participants is the same across different experimental setups. And each task was conducted by three different users with these three different settings, respectively. This can avoid the effect of user's knowledge growth through the prior search session. And no users conducted two tasks in the same topic, which also can avoid the task learning effects on the results. Note that the tasks have negligible or no learning effects on each other even in the same domain if they are not in the same topics. Table 4 shows an example that although the two tasks are all criminal tasks, users cannot increase knowledge about a task after completing another task.

We recruited 15 graduate students from law school (5 for civil law, 5 for criminal law and 5 for commercial law) to be agents. They were all native Chinese speakers and qualified in legal practice.[1] To ensure an adequate level of domain expertise, they only participated in the task related to their research fields. In addition, they all achieved a score of 95 or more in the courses corresponding to their experimental topics. Each task was conducted by three different agents with these three different settings, respectively. This can also avoid the effect of agent's knowledge growth through the prior search session. And they were trained with 5 auxiliary search tasks beforehand to familiarize with the query construction skills in the legal case retrieval system, guaranteeing an adequate level of search expertise.

As for the legal case retrieval system, we choose a leading commercial legal search engine[2] in China. Users and agents had a conversation (just in text form) via Zoom.[3]

[1] They had passed the "National Uniform Legal Profession Qualification Examination".
[2] https://ydzk.chineselaw.com/case
[3] https://zoom.us/

Table 5
The statistics of the dataset in our three groups of user study. #Total indicates the averaged number of cases that were returned by the legal case retrieval system. #Annotated indicates the averaged number of cases clicked by the agent. #Relevant indicates the averaged number of cases regarded as relevant by the users. Knowledge, Interest and Difficulty indicates the average of users' perceived five-grade responses about their domain knowledge, prior interest and task difficulty in the pre-search questionnaires, respectively.

Group	#Total	#Annotated	#Relevant	Knowledge	Interest	Difficulty
Whole	15.45	4.143	1.624	2.404	2.467	3.435
-QG	18.55	5.817	1.431	2.550	2.224	3.326
-BM	16.92	7.211	2.271	2.514	2.601	3.398

Table 6
The interaction effects of the task order, task domain, prior knowledge, task difficulty and prior interest on the experimental outcomes. * indicates that this factor has significant effects on the experimental results at 0.05 level using ANOVA-test.

Measure	Task order			Task domain			Prior knowledge*			Task difficulty*			Prior interest		
	R1	R2	R3	Civil	Criminal	Commercial	Low	Middle	High	Low	Middle	High	Low	Middle	High
Accuracy of visited results	0.712	0.718	0.725	0.733	0.718	0.711	0.672	0.701	0.736	0.754	0.721	0.681	0.712	0.726	0.718
Satisfaction	3.877	3.952	3.901	3.877	3.952	3.901	3.721	3.912	4.212	4.324	3.988	3.561	3.902	3.888	3.952

5.2. Procedure

Before the experiments, we firstly requested each participant to complete a warm-up search task. We then introduce the details of the procedure as follows:

Query Case and Issue Reading. In the first step, the user read the query case description and the legal issue carefully. She could refer to the query case at any time during the session, so she did not need to memorize the case description at this step.

Pre-task Questionnaire. Next, the user was asked to finish a pre-search questionnaire, including: domain knowledge level, task difficulty level, and prior interest level of the task with a 5-point Likert scale (1: not at all, 2: slightly, 3: somewhat, 4: moderately, 5: very).

Task Completion. After that, the user started performing searches with her experimental setting ("Whole", "-QG" or "-BM"). At this step, we collected the agent's interactions with the system, including queries, clicks, etc. Moreover, we recorded the conversation contents, including users' legal questions, agents' clarifying questions, the cases returned by the agent.

Post-task Questionnaire. After examining the supporting cases returned by the agent, the user was required to complete a post-task questionnaire. At this step, we collected explicit feedback signals with respect to the search experience, including five-grade workload and satisfaction.

Result Assessment. After completing the post-task questionnaire, the user was further asked to annotate the cases that agents clicked in the SERPs. That is, a relevance score is annotated to each case (1: irrelevant, 2: relevant). As for the cases that were not clicked, we simply regarded them as irrelevant.

5.3. Data statistics

Table 5 shows the averaged number of cases that were returned by the legal case retrieval system (#Total), clicked by the agent (#Annotated) and regarded as relevant by the users (#Relevant) respectively. In addition, we evaluate users' domain knowledge, prior interest and task difficulty through their perceived five-grade response. The results are also shown in Table 5 (i.e., Knowledge, Interest and Difficulty). We find that there are no significant differences in the three indicators between the "Whole" group and the other two groups at 0.05 level using Mann–Whitney U test. It reveals that individual variability does not make experimental conclusions unreliable.

Then we investigate the interaction effects of task order, task domain, prior knowledge, task difficulty and prior interest on experimental results. Specifically, we choose two indicators: *Accuracy of visited results and Satisfaction*. They denote the precision of the cases that users clicked and their perceived five-grade satisfaction response, respectively. We compare them and perform a series of one-way ANOVA-tests and pairwise t-tests to determine the significance. And the results are shown in Table 6.

As for task order, considering one user conducted three different tasks with different experimental setting, we investigate whether the order in which users completed the three tasks affect the experimental results. We group each task with a certain experimental setting into three categories (R1, R2, R3) in the order of user completion. Here a task with a certain setting in R1, R2 and R3 denotes that a user's first, second and third task, respectively. And we sampled 69 tasks with 69 users for analyzing the interaction effects. Each task was conducted by three different users with three different settings and they satisfy the following conditions: (1) R1, R2, R3 have the same set of users and the same set of tasks (i.e., they all cover 69 users and tasks). (2) The experimental settings in R1, R2 and R3 are all balanced. That is to say, they all contain 23 tasks of the "Whole" group, 23 tasks of the "-QG" group and 23 tasks of the "-BM" group. We find that there is no significant difference in these two indicators between R1, R2 and R3 (ANOVA-$p > 0.05$, $p > 0.05$). It indicates that task order has no significant effect on the experimental results. Because each user was trained with one

auxiliary search tasks to get familiar with the experimental system and the current experimental setting avoids the task learning effects on the results.

And we investigate whether the domain of task affects the experimental results. We sample 18 tasks from each of the three domains and each task was also conducted by three different users with three different settings. And the three domains contain the same set of users. That is to say, the tasks in each domain are done by the same 54 users. We find that there is no significant difference in these two indicators between these three domains (ANOVA-$p > 0.05$, $p > 0.05$). It indicates that task domain has no significant effect on the experimental results.

As for prior knowledge, we group each task with a certain experimental setting into three categories: Low (i.e., 1–2 scores), Middle (i.e., 3 scores) and High (i.e., 4–5 scores), according to the user's perceived five-grade response of prior knowledge. And we sampled 72 tasks with 72 users for analyzing its interaction effects. Each task was conducted by three different users with three different settings and they satisfy the following conditions: (1) Low, Middle and High have the same set of users and the same set of tasks (i.e., they all cover 72 users and tasks). (2) The experimental settings in Low, Middle and High are all balanced. That is to say, they all contain 24 tasks of the "Whole" group, 24 tasks of the "-QG" group and 24 tasks of the "-BM" group. We find that there are significant differences in these two indicators between Low, Middle and High (ANOVA-$p < 0.05$, $p < 0.05$). However, we have shown that there are no significant differences in prior knowledge between the "Whole" group and the other two groups at 0.05 level using Mann–Whitney U test. Therefore, the interaction effects of prior knowledge do not affect the reliability of the final result. Similarly, it is the same for task difficulty and prior interest.

In summary, we acknowledge that there are some assignment bias and environmental factors in current experimental setting. But we illustrate that the experimental results are still convinced based on the above analysis of interaction effects of these factors.

6. Results

6.1. Comparison of agent actions

To address **RQ3**, we compare agent actions in conversational legal case retrieval between while using the complete workflow and removing one of the key modules (i.e., Query Generation or Buffer Mechanism) from the workflow. Following the setting in Section 3.2, we also divide the measures into two groups: Agent-system interaction and Agent-user interaction. The results are shown in Table 7.

6.1.1. Effects of Query Generation

In regard to agent-system interaction, we first investigate whether agents can construct more appropriate queries to better express his understanding of information need with Query Generation. Specifically, we examine the quality of the query in terms of retrieval accuracy over the retrieved cases by three different indicators: *Precision (each turn), Precision (last turn) and Precision (visited results)*. Here, *Precision (each turn)* is the averaged proportion of relevant cases in all the turns of queries which were made by the agent. *Precision (last turn)* is the proportion of relevant cases in the last turn of queries. *Precision (visited results)* is the proportion of relevant cases of the results clicked by the agent. From Table 7, we can see that these three precision scores with Query Generation are higher than those without it. We can also observe that the agents with Query Generation issue less (2.235 vs 3.412) and shorter (2.615 vs 3.523) queries to the systems and click less results (4.143 vs 5.817) than the counterpart without Query Generation. And the agents with Query Generation click more results per query (1.854 vs 1.705). It reveals that the agents with Query Generation are more confident about their queries. Therefore, they tend to continue browsing the SERPs rather than replacing the query. Moreover, the proportion of the clicked items that are further identified by the agent as relevant is much higher when Query Generation is available (16.12% vs 12.35%). It shows that Query Generation can help agents express his understanding of the information need more concise and accurate, which can tackle **Challenge 1** mentioned in Section 3.

As for agent-user interaction, we find that the agents with Query Generation issue less utterances than the counterpart without this module (10.82 vs 13.11), especially asking less clarifying questions (4.109 vs 4.971). And "open-type" clarifying questions compose 57.96% of agents' questions with Query Generation, which is significantly less than the ratio without Query Generation. In addition, the agents with Query Generation can satisfy users' information need by returning less answers (1.301 vs 1.584). It reveals that Query Generation helps agents retrieve relevant cases easier. These illustrate that with Query Generation, agents take less effort to ask clarifying questions and obtain a better expression for the user's information need.

6.1.2. Effects of buffer mechanism

Recall that we introduce a Buffer Mechanism to save all the cases from the SERPs of the queries issued by the agent in a session. As reported in Table 7, there is no significant difference with respect to the query performance in terms of *Precision (each turn)* and *Precision (last turn)* by removing Buffer Mechanism or not. Even the *Precision (last turn)* of agents without Buffer Mechanism is slightly higher because they focus on finding relevant cases in the last turn. However, it is obvious that the proportion of relevant cases among the ones clicked by the agents with Buffer Mechanism is higher than that without Buffer Mechanism (39.2% vs 31.5%). In order to understand this phenomenon in depth, we compare agents' actions in query formulation and result examination. We find that agents with Buffer Mechanism issue 2.235 queries to the system on average, which is significant less than that from agents without Buffer Mechanism (i.e., 2.871 queries). Furthermore, agents with Buffer Mechanism issue shorter queries (2.615 vs 3.891) and click less results per query (1.854 vs 2.511). The proportion of candidates that are used as answers are higher when Buffer Mechanism is in agents' workflow (16.12% vs 10.91%). These indicate that agents without Buffer Mechanism need to formulate

Table 7
Comparison of agent actions. † indicates that the difference between the workflow removing one of the key modules and the complete workflow is statistically significant at 0.05 level using Mann–Whitney U test.

Group	Measure	Whole	-QG	-BM
Agent-system	Precision (each turn)	0.317	0.216†	0.298
	Precision (last turn)	0.433	0.287†	0.436
	Precision (visited results)	0.392	0.246†	0.315†
	#Queries	2.235	3.412†	2.871†
	#Avg words per query	2.615	3.523†	3.891†
	#Clicks avg query	1.854	1.705†	2.511†
	#Clicks	4.143	5.817†	7.211†
	%Candidate->answer	16.12	12.35†	10.91†
Agent-user	#Utterances	10.82	13.11†	11.53
	#Questions	4.109	4.971†	4.251
	%clarify-yes-no	37.82	23.51†	36.04
	%clarify-choice	4.214	4.642	3.982
	%clarify-open	57.96	71.58†	59.98
	#Answer	1.301	1.584†	1.499†

Table 8
Comparison of user behaviors and outcome. † indicates that the difference between the workflow removing one of the key modules and the complete workflow is statistically significant at 0.05 level using Mann–Whitney U test.

Group	Measure	Whole	-QG	-BM
Behavior	#Utterances	10.82	13.11†	11.53
	#Avg words per utterance	20.93	26.72†	21.08
	#Avg words per conversation	113.2	175.1†	121.5
	#Clarifying answer	4.109	4.971†	4.251
	#Avg words per clarifying answer	17.63	24.15†	17.45
	#Avg words of clarifying answer per conversation	72.44	127.3†	77.67
	#Cases per conversation	1.733	2.564†	2.382†
Outcome	Success	0.872	0.734†	0.812†
	Accuracy of visited results	0.783	0.643†	0.703†
	Workload	2.106	3.123†	2.532†
	Satisfaction	4.388	3.671†	3.873†

queries and examine results more carefully because they are not able to revisit the previous results. Furthermore, although most of the interactions between the agents and the users do not change significantly whether Buffer Mechanism is removed or not, agents with Buffer Mechanism satisfy the users' information need with less answers (1.301 vs 1.499). This confirms again that Buffer Mechanism facilitates a legal agent to better distinguish whether a candidate case is relevant, which can tackle **Challenge 2** mentioned in Section 3.

6.1.3. Summary

Regarding **RQ3**, our findings are as follows: (1) As for agent-system interactions, Query Generation can improve the agents' query quality. It can also save agents' effort in formulating queries, examining results and finding relevant cases. While interacting with users, Query Generation can help agents take less effort to ask clarifying questions and satisfy users' information need by answering fewer times. (2) It is easier for agents with Buffer Mechanism to formulate queries and examine results because they are able to revisit the previous results. This helps agents distinguish whether a candidate case is relevant and satisfy the user's information need with less answers.

6.2. Comparison of user behaviors and outcome

To address **RQ4**, we further investigate the impact of Query Generation and Buffer Mechanism towards the user behaviors and their outcome in conversational legal case retrieval. The results are shown in Table 8.

6.2.1. Effects of query generation

As for user interaction behaviors, we find that when Query Generation is available, a user issues less (10.82 vs 13.11) and shorter (20.93 words vs 26.72 words) utterances than those when interacting with an agent without Query Generation. Moreover, they answer less clarifying questions to express their information need when interacting with a conversational agent with Query Generation (4.109 vs 5.271). And the length of clarifying answers is also shorter in this kind of situation than the counterpart without Query Generation (18.73 vs 23.56). This is because agents with Query Generation ask less "open-type" clarifying questions. What is more, users read less cases (1.733 vs 2.564) when searching with a conversational agent with Query Generation in legal case retrieval. These indicate that while interacting with a conversational agent with Query Generation for legal case retrieval, users spend less effort in answering clarifying questions to express their information need and examining cases from the agents.

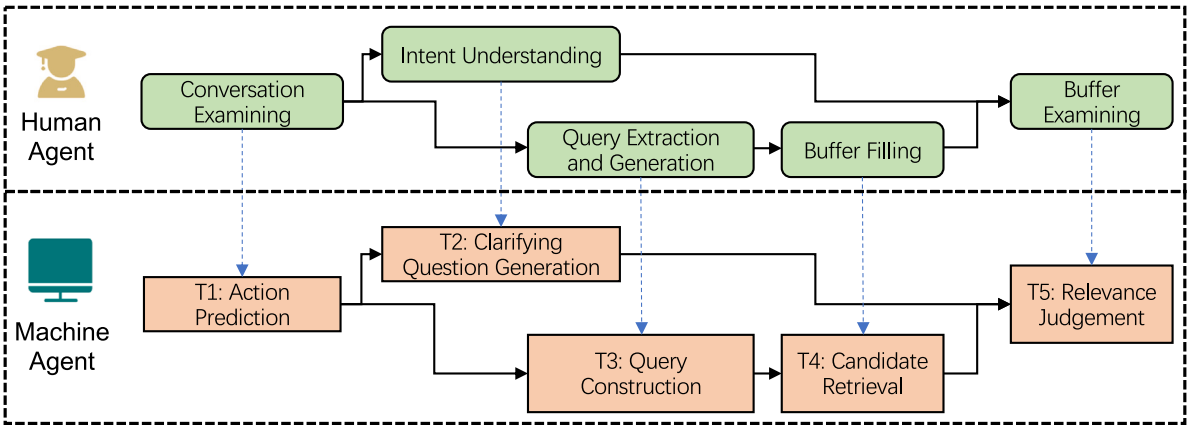

Fig. 3. Five computational tasks to construct a machine agent for conversational legal case retrieval.

Regarding user search outcome, we focus on four indicators: success, accuracy of visited results, workload and satisfaction. Search success measures the objective outcome of a search process (Ageev, Guo, Lagun, & Agichtein, 2011; Odijk, White, Hassan Awadallah, & Dumais, 2015). Specifically, we measure the proportion whether users have found at least one relevant case. Interacting with agents using the complete workflow, users have found at least one relevant case in 87.2% tasks. Interacting with agents without Query Generation, users have found at least one relevant case in 73.4% tasks. This illustrates Query Generation can improve search success in legal case retrieval. In addition, we compare *accuracy of visited results*, which represents the precision of the cases that users clicked, to measure search accuracy. It is obvious that search accuracy improves significantly when agents' workflow contains Query Generation. At last, we evaluate users' workload and satisfaction (Kelly, 2009) through their perceived five-grade response. Users reported 2.106 workload scores and 4.388 satisfaction scores when searching using the complete workflow, respectively. In comparison, users reported 3.123 workload scores and 3.671 satisfaction scores when searching without Query Generation. These indicate that Query Generation helps users spend less effort and achieve higher satisfaction in legal case retrieval.

6.2.2. *Effects of Buffer Mechanism*

We find that most user interaction behaviors do not change significantly whether the agents' workflow contains Buffer Mechanism. However, when Buffer Mechanism is included, a user on average reads 1.733 cases in a session, which is significant less than without Buffer Mechanism. In addition, although many measures (i.e., the number of utterances, the number of clarifying questions and the proportion of each type of clarifying questions) with respect to the interactions between a legal agent and a user do not change significantly when Buffer Mechanism is included or not, Buffer Mechanism improves search accuracy by returning less cases. And Buffer Mechanism achieves higher search success because agents with it return relevant cases by less times and avoids that users' patience are exhausted. Also, when we do not use Buffer Mechanism, the users reported a workload score of 2.532 and a satisfaction score of 3.873, which are significantly worse. In other words, the user spends more efforts and perceives lower satisfaction when the agents do not use Buffer Mechanism.

6.2.3. *Summary*

In summary, our findings towards answering **RQ4** are as follows: (1) Query Generation helps users spend less effort in answering clarifying questions to express their information needs and examining cases in conversational legal case retrieval. (2) Buffer Mechanism makes users read less cases. (3) Both the modules help users save efforts and achieve higher satisfaction.

7. Practical implications

In this section, we the practical implications of this research. Specifically, in our proposed workflow (as shown in Fig. 2), there are five kinds of agent actions, namely conversation examining, intent understanding, query extraction and generation, buffer filling and buffer examining. Therefore, we formalize five computational tasks corresponding to the five kinds of actions to construct a machine agent for conversational legal case retrieval, which are shown in Fig. 3.

T1: Action Prediction. To connect users to the right information with maximal utility, a good conversational legal case retrieval system should be able to take appropriate actions at the right time, i.e., helping the users to clarify their intents whenever unclear, and submitting queries to the system to collect candidate cases. This task has been well studied in task-oriented dialogue systems (Peng et al., 2017) and conversational systems for general web search (Wang & Ai, 2021). However, it is more difficult for agents to understand users' intents in legal case retrieval so that we should design a model that is more inclined to ask clarifying questions.

T2: Clarifying Question Generation. Asking clarifying questions is particularly important for conversational search systems (Aliannejadi, Zamani, Crestani, & Croft, 2019) and several frameworks (Zamani et al., 2020) have been proposed to solve

Table 9
The statistics of the large-scale dataset. #Query and #Unique Generated Query Term both correspond to queries made by the agents submitted to the system.

#Task	#Clarifying question	#Query	#Unique generated query Term	#Candidate	#Relevant
1556	6309	3481	5911	10739	1982

it. In legal case retrieval, we need to propose a new taxonomy of clarifying questions instead of simply dividing them into three categories: yes–no, choice and open. Based on the taxonomy, we can summarize the characteristic of the agent's clarifying questions and design the corresponding model for legal case retrieval.

T3: Query Construction. Ren et al. (2021) define this task as extracting keyphrases from the conversation in general web search. And our results reveal that simply extracting keyphrases is not sufficient to precisely express the information need for agents in legal case retrieval. Therefore, we need to design a model that combines keyphrase extraction and generation to help traditional search engines retrieve more relevant candidate cases (corresponding the **query generation module**).

T4: Candidate Retrieval. Considering it is more difficult for agents to distinguish whether a candidate case should be chosen exactly, we choose top-k candidates of query construction results and utilize all of them to retrieve candidates (corresponding the **buffer mechanism module**). This can help our models to retrieve candidates related to different charges and compare their fine-grained information.

T5: Relevance Judgement. There are several frameworks aiming to solve legal case retrieval task and most of them calculate the relevant score of two cases (Ma et al., 2021; Shao, Liu, Mao, Liu, Zhang & Ma, 2020; Shao, Mao et al., 2020). And in conversational legal case retrieval, models for relevance judgement need to compare retrieved cases from different queries.

Following these five tasks, we further build a large-scale dataset for conversational legal case retrieval. Specifically, we collected 1556 search tasks from legal practitioners' real information need via online forums and social networks. And we recruited the same human agents as our user study and 50 participants to be users who also have participated in our user study. The agents completed these 1556 search tasks following the workflow in our developed experimental platform. Unlike the previous user study, the agents need to label their actions explicitly in the search process, which helps provide labels for training models. Specifically, as for **T1**, the agents explicitly choose to ask clarifying questions or collect candidates. As for **T3** and **T4**, the agents explicitly choose the number of queries and label the query terms which are generated rather than extracted. Table 9 shows the statistics of this dataset.

This dataset can be used for model training and evaluation. We tried two basic tasks on this dataset: whether to use Query Generation now and whether to use Buffer Mechanism now. The first one denotes that based on the current conversation, the agents should generate query terms by themselves or just extract query terms from the conversation. The second one denotes that based on the current conversation, the agents should construct single or multiple queries to retrieve candidates. We simply fine-tune Lawformer (Xiao, Hu, Liu, Tu, & Sun, 2021) on our dataset. Here LawFormer is a Longformer-based pre-trained language model for Chinese legal long documents understanding. F1-scores of these two tasks can achieve 0.947 and 0.968, respectively. This shows that our dataset can train the model well and is ready for the implementation of these two key modules.

Furthermore, we discuss about how to implement Query Generation and Buffer Mechanism in machine agents. In terms of Query Generation, we summarize the following guidance:

- **How to build a legalese database to provide candidates for generating query terms?** Following the dataset construction process of LEVEN (Yao et al., 2022) (a dataset of Chinese legal event detection), we recommend utilizing law articles and classical legal textbooks legal professional references to collect causes and charges which are high-level summaries of legal issues in cases.
- **How to effectively combine extraction and generation?** The form of this problem is the same as that of Out-Of-Vocabulary (See, Liu, & Manning, 2017) (OOV) problem. Specially, See et al. (2017) learn a probability pointer P, which is used to determine whether to generate or copy (namely extraction) at the current time step.

In terms of Buffer Mechanism, we summarize the following guidance:

- **How to select multiple charges related to the conversation for constructing queries?** Considering Buffer Mechanism makes agents compare cases from different queries related to a variety of possible charges, how to select charges has a great impact on the quality of candidate cases. We suggest two ways: (1) simply apply Sentence-Bert (Reimers & Gurevych, 2019) to match charges and the conversation; (2) regard the charges as sub-topics and utilize diverse search models (Feng, Xu, Lan, Guo, Zeng, & Cheng, 2018).
- **How to rerank the cases from different queries?** Existing methods for legal case retrieval aims to match two semi-structured documents, which are not suitable for conversational legal case retrieval. And the candidate cases has covered a variety of possible charges related to the conversation with the help of buffer. Therefore, we recommend applying event detection models (Wang, Han, Liu, Sun, & Li, 2019) to get the event types in the conversation and candidate cases. And matching the event types in the conversation and candidate cases can help improve the rerank performance.

8. Conclusion

In this paper, we focus on how to construct a suitable conversational agent workflow for legal case retrieval. We first conducted a preliminary study to investigate the differences of search process between legal case retrieval and general web search in conversational search paradigm. And we concluded that the proposed conversational agent workflow for general web search faces two serious challenges when applied in the scenario of legal case retrieval: (1) It is more complex for the agent to express his understanding of users' information need by organizing queries. (2) It is more difficult for agents to distinguish whether a candidate case should be chosen exactly.

To tackle the challenges, we propose a suitable conversational agent workflow for legal case retrieval, which contains two additional key modules: **Query Generation** and **Buffer Mechanism**. And we investigate whether the two modules can be adopted to help the conversational agents work better and improve the users' legal case retrieval experience. Centered on the research questions, we conducted three groups of user study and have obtained several interesting findings. (1) Agents can ask less clarifying questions and improve query quality with Query Generation. Therefore, users can save efforts in answering clarifying questions. (2) Buffer Mechanism can save agents' efforts in examining candidate cases, leading to a more precise answer result. (3) Both of the modules can help users save efforts and achieve higher satisfaction. Our results reveal the effectiveness of our proposed conversational agent workflow for legal case retrieval. As for future work, we will design models for conversational legal case retrieval based on our dataset following the workflow.

We acknowledge that there are some limitations of our user study design. On the one hand, although we show that factors such as individual variability did not significantly change the experimental results under the current experimental setup, the effects of these factors cannot be completely eliminated. On the other hand, the gender of the participants are not balanced due to the low percentage of male students in the law school and the large number of participants recruited in this experiment.

CRediT authorship contribution statement

Bulou Liu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Yueyue Wu:** Conceptualization, Methodology, Data curation. **Fan Zhang:** Conceptualization, Methodology, Writing – review & editing. **Yiqun Liu:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Zhihong Wang:** Methodology, Writing – review & editing. **Chenliang Li:** Methodology, Writing – original draft. **Min Zhang:** Funding acquisition. **Shaoping Ma:** Funding acquisition.

Data availability

Data will be made available on request.

References

Ageev, M., Guo, Q., Lagun, D., & Agichtein, E. (2011). Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on research and development in information retrieval* (pp. 345–354).

Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International acm sigir conference on research and development in information retrieval* (pp. 475–484).

Arewa, O. B. (2006). Open access in a closed universe: Lexis, westlaw, law schools, and the legal information market. *The Lewis & Clark Law Review*, *10*, 797.

Azzopardi, L., Dubiel, M., Halvey, M., & Dalton, J. (2018). Conceptualizing agent-human interactions during the conversational search process. In *The second international workshop on conversational approaches to information retrieval*.

Balog, K., Flekova, L., Hagen, M., Jones, R., Potthast, M., Radlinski, F., et al. (2020). Common conversational community prototype: Scholarly conversational assistant. CoRR arXiv:2001.06910.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, *9*(3), 379–395.

Carterette, B., Kanoulas, E., Hall, M. M., & Clough, P. D. (2014). Overview of the TREC 2014 session track. In *NIST Special Publication, Proceedings of the twenty-third text retrieval conference, TREC 2014*.

Doyle, J. (1992). WESTLAW and the American digest classification scheme. *Law Library Journal*, *84*, 229.

Feng, Y., Xu, J., Lan, Y., Guo, J., Zeng, W., & Cheng, X. (2018). From greedy selection to exploratory decision-making: Diverse ranking with policy-value networks. In *The 41st International ACM SIGIR Conference on research & development in information retrieval* (pp. 125–134).

Ferrer, A. S., Hernández, C. F., & Boulat, P. (2014). LEGAL SEARCH: foundations, evolution and next challenges. The wolters kluwer experience. *Revista Democracia Digital E Governo EletrÔNico*, *1*, 120–132.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Hamann, H. (2019). The german federal courts dataset 1950–2019: From paper archives to linked open data. *Journal of Empirical Legal Studies*, *16*(3), 671–688.

Hashemi, H., Zamani, H., & Croft, W. B. (2020). Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd International Acm Sigir Conference on research and development in information retrieval* (pp. 1131–1140).

Jackson, P., Al-Kofahi, K., Tyrrell, A., & Vachher, A. (2003). Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, *150*(1–2), 239–290.

Kelly, D. (2009). *Methods for evaluating interactive information retrieval systems with users*. Now Publishers Inc.

Liu, B., Wu, Y., Liu, Y., Zhang, F., Shao, Y., Li, C., et al. (2021). Conversational vs traditional: Comparing search behavior and outcome in legal case retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on research and development in information retrieval* (pp. 1622–1626).

Ma, Y., Shao, Y., Liu, B., Liu, Y., Zhang, M., & Ma, S. (2021). Retrieving legal cases from a large-scale candidate corpus. In *Proceedings of the 18th International conference on artificial intelligence and law*.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*, 50–60.

Mao, J., Liu, Y., Kando, N., Zhang, M., & Ma, S. (2018). How does domain expertise affect users' search interaction and outcome in exploratory search? *ACM Transactions on Information Systems (TOIS)*, *36*(4), 1–30.

McGinnis, J. O., & Pearce, R. G. (2019). The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Actual Problems of Economics and Law*, *13*, 1230.

McGinnis, J. O., & Wasick, S. (2014). Law's algorithm. *Florida Law Review*, *66*, 991.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., et al. (2016). MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*.

Odijk, D., White, R. W., Hassan Awadallah, A., & Dumais, S. T. (2015). Struggling and success in web search. In *Proceedings of the 24th ACM International on conference on information and knowledge management* (pp. 1551–1560).

Peng, B., Li, X., Li, L., Gao, J., Celikyilmaz, A., Lee, S., et al. (2017). Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on empirical methods in natural language processing* (pp. 2231–2240).

Radlinski, F., & Craswell, N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on conference human information interaction and retrieval* (pp. 117–126).

Reimers, N., & Gurevych, I. (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, (pp. 3982–3992).

Ren, P., Liu, Z., Song, X., Tian, H., Chen, Z., Ren, Z., et al. (2021). Wizard of search engine: Access to information through conversations with search engines. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval* (pp. 533–543).

Rose, D. E., & Belew, R. K. (1989). Legal information retrieval a hybrid approach. In *Proceedings of the 2nd International conference on artificial intelligence and law* (pp. 138–146).

Saravanan, M., Ravindran, B., & Raman, S. (2009). Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, *17*(2), 101–124.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Shao, Y., Liu, B., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2020). THUIR@ COLIEE-2020: Leveraging semantic understanding and exact matching for legal case retrieval and entailment. arXiv preprint arXiv:2012.13102.

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., et al. (2020). BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI* (pp. 3501–3507).

Shao, Y., Wu, Y., Liu, Y., Mao, J., Zhang, M., & Ma, S. (2021). Investigating user behavior in legal case retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on research and development in information retrieval* (pp. 962–972).

Solomon, P. (1997). Conversation in information-seeking contexts: A test of an analytical framework. *Library & Information Science Research*, *19*(3), 217–248.

Thomas, P., McDuff, D., Czerwinski, M., & Craswell, N. (2017). Misc: A data set of information-seeking conversations. In *Proceedings of the 1st International workshop on conversational approaches to information retrieval*.

Trippas, J. R., Spina, D., Cavedon, L., Joho, H., & Sanderson, M. (2018). Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on human information interaction & retrieval* (pp. 32–41).

Turtle, H. (1995). Text retrieval in the legal world. *Artificial Intelligence and Law*, *3*(1), 5–54.

Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, *25*(1), 65–87.

Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. (2017). Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 Chi Conference extended abstracts on human factors in computing systems* (pp. 2187–2193).

Wang, Z., & Ai, Q. (2021). Controlling the risk of conversational search via reinforcement learning. In *Proceedings of the web conference 2021* (pp. 1968–1977).

Wang, X., Han, X., Liu, Z., Sun, M., & Li, P. (2019). Adversarial training for weakly supervised event detection. 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)* (pp. 998–1008).

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, *2*, 79–84.

Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., et al. (2022). LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for computational linguistics: ACL 2022* (pp. 183–201).

Yu, S., Liu, Z., Xiong, C., Feng, T., & Liu, Z. (2021). Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval* (pp. 829–838).

Zamani, H., & Craswell, N. (2020). Macaw: An extensible conversational information seeking platform. In *Proceedings of the 43rd International ACM SIGIR Conference on research and development in information retrieval* (pp. 2193–2196).

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020* (pp. 418–428).

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th Acm International conference on information and knowledge management* (pp. 177–186).