Received 4 June 2024
Revised 28 November 2024
2 February 2025
Accepted 20 March 2025

Building a construction law knowledge repository to enhance general-purpose large language model performance on domain question-answering: a case of China

Shenghua Zhou
China-Pakistan Belt and Road Joint Laboratory on Smart Disaster Prevention of Major Infrastructures, Southeast University, Nanjing, China and Department of Engineering, University of Cambridge, Cambridge, UK
Hongyu Wang
China-Pakistan Belt and Road Joint Laboratory on Smart Disaster Prevention of Major Infrastructures, Southeast University, Nanjing, China
S. Thomas Ng
Department of Architecture and Civil Engineering, City University of Hong Kong, Kowloon, Hong Kong
Dezhi Li and Shenming Xie
China-Pakistan Belt and Road Joint Laboratory on Smart Disaster Prevention of Major Infrastructures, Southeast University, Nanjing, China
Kaiwen Chen
The University of Alabama System, Tuscaloosa, Alabama, USA, and
Wentao Wang
China-Pakistan Belt and Road Joint Laboratory on Smart Disaster Prevention of Major Infrastructures, Southeast University, Nanjing, China

Abstract

Purpose – Achieving smart question-answering (QA) for construction laws (CLs) holds significant promise in aiding domain professionals with legal inquiries. Existing studies of construction law question-answering (CLQA) rely on learning-based models, which require extensive training data and are limited to a narrow QA scope. Meanwhile, general-purpose large language models (GPLLMs) possess great potential for CLQA but fall short of domain-specific knowledge. This study aims to propose a data-driven and expertise-based approach to develop a construction law knowledge repository (CLKR) and validate its effectiveness in enhancing the CLQA performance of GPLLMs.

Design/methodology/approach – This methodology includes (1) recognizing 702 candidate CL documents from 374,992 official judgments, (2) building a CLKR with 387 filtered documents covering eight CL knowledge areas, (3) integrating CLKR and seven representative GPLLMs and (4) constructing a 2,140-question CLQA dataset from Professional Construction Engineer Qualification Examinations (PCEQEs) during 2014–2023 to compare CLQA performance between seven pairs of GPLLMs with and without CLKR.

Engineering, Construction and Architectural Management
Vol. 32 No. 13, 2025
pp. 518-546
Emerald Publishing Limited
e-ISSN: 1365-232X
p-ISSN: 0969-9988
DOI 10.1108/ECAM-06-2024-0701

This study is financially supported by National Natural Science Foundation of China (No. 72201057) and Social Science Foundation of Jiangsu Province (No. 23GLC020). All data, codes, results, and videos are provided in the supplemental materials in the GitHub repository (https://github.com/0AnonymousSite0/Question_Answering_of_Construction_Laws).

Findings – The CLKR significantly enhances the CLQA performance of seven GPLLMs, yielding an impressive average accuracy increase of 21.1%, with individual improvements ranging from 9.9 to 44.9%. Furthermore, CLKR boosts the accuracy of single-answer questions by 14.9% and multiple-answer questions by 38.3%. Additionally, the accuracy enhancements across 8 CL knowledge areas are between 14.5 and 28.2%.
Originality/value – This study proposes an approach of developing the external knowledge base of CLKR to empower GPLLMs, significantly expanding the scope of CLQA while bypassing the complex training of traditional learning-based models. Moreover, this study confirms the effectiveness of CLKR in augmenting GPLLM performance and offers a reusable CLQA test dataset as a benchmark.

Keywords Large language models, Question-answering, Construction laws, Knowledge repository
Paper type Research paper

1. Introduction

The broadly defined construction laws (CLs) include various statutes, acts, regulations, judicial decisions, and legal interpretations (Alhyari and Ani, 2022; Khademi Adel *et al.*, 2022). These construction laws influence multiple critical aspects of the construction industry, such as contractual breaches (Khademi Adel *et al.*, 2022), safety incident treatment (Zailani *et al.*, 2023), occupational health risks (Tao *et al.*, 2022; Liu *et al.*, 2021), and others. Addressing such legal issues within the construction industry is crucial for the smooth implementation of construction projects (Tao *et al.*, 2022; Zailani *et al.*, 2023). The question-answering (QA) for CLs currently relies on three main approaches: (1) consulting literature, (2) searching for online information, or (3) seeking guidance from domain experts. Reviewing literature (e.g. books, standards, and statutes) is a laborious process, often requiring the review of hundreds of documents to answer a query (Rasool *et al.*, 2024; Choi *et al.*, 2023). Search engines are freely accessible, but they only provide information relevant to the question, instead of directly delivering specific answers (Oeding *et al.*, 2024; Liu *et al.*, 2024). Consulting domain experts is an effective approach, but the scarcity of experts results in high costs (Alan *et al.*, 2024; Hou and Zhang, 2024). For example, only 8.24 million out of 33.72 million (24.44%) legal cases in Mainland China involved lawyers until 2023, leaving 75.56% of cases without lawyers (DJ, 2023; SPC, 2023). Therefore, global academia and industry face the challenge of efficiently and cost-effectively providing CLQA. The smart CLQA that automatically answers construction-related legal questions would be a valuable complement, helping resolve the basic legal queries in the construction domain with reduced cost and time consumption.

General-purpose large language models (GPLLMs) (e.g. ChatGPT and Llama) are large language models designed to handle a broad range of language tasks across multiple domains without specific domain specialization (Rizzo *et al.*, 2024; Gilson *et al.*, 2023). GPLLM-based CLQA is advantageous for being training-free and having a strong language understanding (Ghimire *et al.*, 2024; Pursnani *et al.*, 2023; Oh *et al.*, 2023). However, these non-specialized GPLLMs lack CL knowledge incorporation, which may result in unsatisfying CLQA performance (Su *et al.*, 2024; Tsoutsanis and Tsoutsanis, 2024; Frieder *et al.*, 2023). Developing external knowledge bases encounters significant challenges: (1) a heavy reliance on expertise-based selection of knowledge-embedded documents, which restricts the CLKR scope and limits updates; (2) a lack of empirical validation regarding whether domain knowledge can enhance CLQA performance; and (3) a shortage of benchmark datasets for CLQA testing.

In response, this study aims to (1) propose a data-driven and expertise-based approach to developing CLKR, (2) validate the effectiveness of CLKR in improving GPLLM performance, and (3) provide an openly available CLQA dataset. To achieve these objectives, this study devises a four-phase CLKR development approach to construct a 387-document CLKR, based on 374,992 written judgments and 10 experts. Besides, this study validates the effectiveness of the CLKR by comparing the performance of seven GPLLMs with and without CLKR, as well as providing a 2,140-question CLQA dataset based on China's most authoritative Professional Construction Engineer Qualification Examinations (PCEQEs) in architecture, engineering, and construction (AEC) field. The remaining sections of this study are organized as follows. Section 2 provides a literature review on existing construction-related QA research and the application of GPLLMs. Section 3 outlines the methodology for building the CLKR and integrating it with

GPLLMs. Section 4 evaluates the effectiveness of the CLKR in enhancing the CLQA of GPLLMs. Section 5 discusses the research contributions, potential methods for the long-tail effect, mis-answered question types, and practical implications.

2. Related works

2.1 Existing construction-focused QA studies

In the construction domain, there are pioneering smart QA studies, as listed in Table 1. Existing construction-focused QA studies involve the procurement of construction materials (Lee *et al.*, 2023), construction safety hazards (Wang and El-Gohary, 2023; Tian *et al.*, 2023), construction accidents (Xu *et al.*, 2023), building codes (Xue *et al.*, 2024), construction document text mining (Sun *et al.*, 2020), and construction procedures (Zhong *et al.*, 2020). Although these studies focus on different topics, they still offer significant insights into (1) exploited QA models and (2) QA performance testing for this study.

Existing construction-related smart QA is primarily constructed upon machine learning and deep learning models (Table 1). Conventional CLQA-used machine learning models are often based on algorithms like TF-IDF (Sun *et al.*, 2020) and logistic regression (Rajpurkar *et al.*, 2016). While these machine learning models could be developed for QA with small amounts of training data, their natural language processing (NLP) capabilities are limited, resulting in suboptimal QA performance (Lee *et al.*, 2023). Deep learning models such as BERT (Chou *et al.*, 2024; Tian *et al.*, 2023; Kim *et al.*, 2024), CNN (Lee *et al.*, 2023; Wang and El-Gohary, 2023), and BiLSTM (Xu *et al.*, 2023; Zhong *et al.*, 2020) exhibit better NLP capabilities and more optimal QA performance, owing to their greater number of parameters. However, the deep learning models necessitate a large volume of data for QA training (Chou *et al.*, 2024; Xue *et al.*, 2024). Annotating QA training data is a high-cost and labor-intensive process (Chou *et al.*, 2024; Lee *et al.*, 2023; Wang and El-Gohary, 2023; Xue *et al.*, 2024). Moreover, whether employing machine learning or deep learning models for QA, a fundamental limitation persists in that these learning-based models are only applicable to questions covered by training data. In other words, their knowledge scope for QA is relatively narrow, such as one procurement document (Lee *et al.*, 2023), two chapters of a construction code (Xue *et al.*, 2024), three standard contracts (Kim *et al.*, 2024), and a dozen or so regulations (Zhong *et al.*, 2020) (Table 1).

In construction-related QA performance testing, QA performance test sets in existing studies have two sources: (1) expert-designed question-answer pairs and (2) test datasets built on authoritative exams (Table 1). The first source relies primarily on domain experts to devise question-answer pairs tailored to specific CL subareas, drawing from relevant literature, project materials, and their expertise (Chou *et al.*, 2024; Lee *et al.*, 2023; Sun *et al.*, 2020). This QA performance test data source often requires a significant investment of human resources and time, resulting in relatively small test datasets with fewer than 671 questions (Table 1) (Xu *et al.*, 2023; Tian *et al.*, 2023). Such limited test datasets may be inadequate for evaluating the QA performance of GPLLMs across a broad domain like construction laws (Martinez-Gil, 2023; Sun *et al.*, 2020). Moreover, due to individuals' knowledge limitations and biases, the authority of expertise-based QA test sets is also often questioned (Zhong *et al.*, 2020). The second source leverages questions from professional examinations (Gencer and Aydin, 2023; Pursnani *et al.*, 2023; Wang and El-Gohary, 2023), which are highly authoritative and facilitate the rapid creation of performance test datasets (Rizzo *et al.*, 2024; Sahin *et al.*, 2024; Frieder *et al.*, 2023). Hence, these examination-based QA test datasets offer valuable guidance for developing the CLKR performance test dataset in this study.

2.2 Studies based on GPLLMs

GPLLMs demonstrate significant advancements in language comprehension, context understanding, and text generation over traditional learning-based models (e.g. machine

Table 1. Existing construction-related QA studies

No	References	QA targeted specific areas	QA-used models	Training-free	Knowledge scope for QA	QA performance test dataset	
						Question source	Number of questions
1	Chou *et al.* (2024)	Risk management in river dredging projects	A BERT-based deep learning model	×	Dredging risk knowledge collected by interviews	Listed by experienced dredging personnel	16
2	Kim *et al.* (2024)	Construction market knowledge in overseas projects	A BERT-based deep learning model	×	3 versions of a FIDIC standard contract written in English, Korean, and Indonesian	The FIDIC documents	80
3	Xue *et al.* (2024)	Building codes	A BERT-based deep learning model	×	2 Chapters of the IBC 2015	Manually generated for model testing	175
4	Lee *et al.* (2023)	Steel manufacturer equipment procurement	A machine learning model combining KG and QA	×	An equipment procurement document from a steel-making company	Generated questions based on relevant arbitration and clause settings	45
5	Tian *et al.* (2023)	Construction safety hazard	A BERT + BiGRU + Self-Attention-based deep learning model	×	6,325 safety hazard texts	Dedicated questions for model application	25
6	Wang and El-Gohary (2023)	Construction safety hazard	A CNN-based deep learning model	×	20 OSHA sections related to fall protection	Manually developed for model testing	671
7	Xu *et al.* (2023)	Coal mine construction safety	A BERT-BiLSTM-CRF-based deep learning model	×	43 sections of 80 papers from coal mine construction safety management standard specifications	Example questions used to validate the semantic query and entity information modules	Unspecified
8	Sun *et al.* (2020)	Construction document information transmission mining	A TF-IDF-based machine learning model	×	A monthly construction report containing 1734 words	Posed by three construction managers	5
9	Zhong *et al.* (2020)	Construction procedural constraint	A BiLSTM- + CRF-based deep learning model	×	14 types of national standards of CACQ in China	Sentences labeled by experts	400
10	Rajpurkar *et al.* (2016)	Multiple domains including building regulation domain	A logistic regression-based machine learning model	×	536 Wikipedia articles	Contributed by 5 civil engineers	Unspecified

Note(s): BERT: Bidirectional Encoder Representations from Transformers; BiGRU: Bidirectional Gated Recurrent Unit; BIM: Building Information Modeling; BiLSTM: Bidirectional Long Short-Term Memory; CACQ: Code for Acceptance of Construction Quality; CRF: Conditional Random Fields; FIDIC: International Federation of Consulting Engineers; IBC: International Building Code; IE: Information Extraction; KG: Knowledge Graph; NHC: National Hurricane Center; NLG: Natural Language Generation; NLP: Natural Language Process; NLU: Natural Language Understanding; OSHA: Occupational Safety and Health Organization; TF-IDF: Term Frequency-Inverse Document Frequency
Source(s): Authors' own work

learning and deep learning models) (Rizzo *et al.*, 2024; Sahin *et al.*, 2024; Oh *et al.*, 2023). These advancements benefit from the remarkable parameter volumes of GPLLMs (Rizzo *et al.*, 2024; Gilson *et al.*, 2023). For example, GPT-4.0 has 1.8 trillion parameters, Llama-2-70b has 70 billion parameters, and ERNIE-Bot-turbo also exceeds 1 trillion parameters. In contrast, conventional learning-based models have much smaller parameter scales; even the largest, such as BERT, only has 340 million parameters (Shi *et al.*, 2023), leading to limited capabilities in language processing (Gilson *et al.*, 2023; Saad *et al.*, 2023). State-of-the-art research has applied GPLLMs in various domains, such as orthopedics (Rizzo *et al.*, 2024; Saad *et al.*, 2023), neurosurgery (Sahin *et al.*, 2024), urology (Schoch *et al.*, 2024), nursing (Su *et al.*, 2024), clinical help (Tsoutsanis and Tsoutsanis, 2024), ophthalmology (Antaki *et al.*, 2023), thoracic surgery (Gencer and Aydin, 2023), mathematics (Frieder *et al.*, 2023), and physics (Kortemeyer, 2023).

Although GPLLMs possess more powerful language processing capabilities, their QA performance in specific professional domains remains inadequate due to a lack of domain knowledge (Lu *et al.*, 2024; Su *et al.*, 2024; Tsoutsanis and Tsoutsanis, 2024). The integration of external knowledge bases is considered to potentially further enhance the performance of GPLLMs, as suggested by recent studies (Lu *et al.*, 2024; Su *et al.*, 2024; Tsoutsanis and Tsoutsanis, 2024). However, few scholars in the AEC field have attempted to develop domain-specific knowledge repositories to lift GPLLMs' QA performance (Ghimire *et al.*, 2024; Antaki *et al.*, 2023) (Table 2). There is still a lack of empirical validation regarding whether domain knowledge could lift the CLQA performance of GPLLMs.

The development of external knowledge bases for the Retrieval-Augmented Generation (RAG) of GPLLMs encounters significant challenges: (1) heavy reliance on expert knowledge to construct the knowledge base, (2) a limited scale of the knowledge base, and (3) the infrequency of updates to the knowledge base. Specifically, the process of building these knowledge bases heavily depends on experts manually selecting domain-relevant documents (Rasool *et al.*, 2024; Alan *et al.*, 2024). This approach risks critical omissions when constructing a large-scale domain knowledge base (e.g. CLKR), thus reducing its comprehensiveness (Lee *et al.*, 2023; Zhong *et al.*, 2020). Furthermore, domain knowledge, such as in CLKR, continually evolves (Ghimire *et al.*, 2024; Khademi Adel *et al.*, 2022). The failure to update the knowledge base can render it not only ineffective but also potentially harmful to GPLLM performance (Gao *et al.*, 2023). Previous research has rarely addressed the issues of updating these knowledge bases or developing their update mechanisms (Hou and Zhang, 2024; Alan *et al.*, 2024; Mansurova *et al.*, 2024).

The exploitation of knowledge bases relies on RAG technology. A variety of RAG frameworks are available, including LangChain (Langchain-ai, 2024), Haystack (deepset-ai, 2024), RAGFlow (infiniflow, 2024), Txtai (neuml, 2024), LLM-App (pathwaycom, 2024), FlashRAG (RUC-NLPIR, 2024), and Cognita (truefoundry, 2024). Each of these RAG frameworks supports a modular architecture with data connectors, document loaders, vector stores, and GPLLM integrations (Hou and Zhang, 2024; Rasool *et al.*, 2024; Mansurova *et al.*, 2024; Petrus, 2024). Among these, LangChain stands out for its open-source nature, active community, and user-friendly API (Alan *et al.*, 2024; Mansurova *et al.*, 2024). Consequently, LangChain is selected for subsequent research to implement the integration between LLMs and external knowledge bases.

The evaluation of QA performance for GPLLMs necessitates test datasets, with question types potentially being open-ended or closed-ended (Table 2). Open-ended questions are typically designed by a few experts and lack authoritative sources (Alan *et al.*, 2024; Choi *et al.*, 2023). These open-ended questions need subjective expert judgment for assessing answer quality, making them unsuitable as benchmark datasets (Harvel *et al.*, 2024; Zheng *et al.*, 2023). In contrast, most closed-ended question sets are based on officially organized exams with objective and consistent answers (Rizzo *et al.*, 2024; Pursnani *et al.*, 2023; Rosól *et al.*, 2023), making them suitable for use as benchmark datasets (Su *et al.*, 2024). Generally, closed-ended questions outperform open-ended ones in terms of question quantity, source

Table 2. Examples of recent studies on GPLLMs

No	References	GPLLMs	Specific domain	Test datasets		
				Question sources	Question types	Number of questions
1	Alan *et al.* (2024)	GPT-3.5 turbo	Islam understanding	Designed by experts	Open-ended	3 (mentioned by the author)
2	Hou and Zhang (2024)	GPT-3.5 and GPT-4.0	Dietary supplement	Information on the MSKCC website	Closed-ended (MSQs and True/False)	2000
3	Mansurova *et al.* (2024)	Llama-2-7b and Llama-2-13b	General	TriviaQA open-domain dataset	Closed-ended (Filling in the blank)	500
4	Rasool *et al.* (2024)	GPT-3.5-turbo and GPT-4	Healthcare	CogTale dataset	Closed-ended (MMQs, MSQs, True/False, and number extraction)	337
5	Rizzo *et al.* (2024)	GPT-3.5 turbo and GPT-4	Orthopaedics	OITE in the 2020, 2021, and 2022	Closed-ended (MSQs)	207
6	Sahin *et al.* (2024)	GPT-4	Neurosurgery	The latest six written TNSPBE	Closed-ended (MSQs)	523
7	Schoch *et al.* (2024)	GPT-3.5 and GPT-4	Urology	A test book published by the FEBU association	Closed-ended (MSQs)	Around 600
8	Su *et al.* (2024)	GPT-4	Nursing	Taiwan's 2022 Nursing Licensing Exam	Closed-ended (MSQs)	400
9	Tsoutsanis and Tsoutsanis (2024)	Llama-2, Google Bard, Bing Chat, and GPT-3.5	Clinical help	Commercial question banks (i.e. Qbank) for the MSRA exam	Closed-ended (MSQs)	100
10	Antaki *et al.* (2023)	GPT-3.5 turbo and GPT-4	Ophthalmology	Basic and Clinical Science Course Self-Assessment Program and an online question bank (i.e. OphthoQuestions)	Closed-ended (MSQs)	520
11	Choi *et al.* (2023)	ChatGPT	Laws	Exams for law school courses at the University of Minnesota	Closed-ended (MSQs) and open-ended (essay writing)	107
12	Gencer and Aydin (2023)	GPT-3.5 and GPT-4	Thoracic surgery	Turkish-language thoracic surgery exam questions	Closed-ended (MSQs)	105
13	Gilson *et al.* (2023)	InstructGPT, GPT-3.5, and ChatGPT	Medicine	A question bank for medical students and the NBME	Closed-ended (MSQs)	220
14	Oh *et al.* (2023)	GPT-3.5 and GPT-4	Surgery	The KGSBE in 2020, 2021, and 2022	Closed-ended (MSQs)	280
15	Pursnani *et al.* (2023)	GPT-3.5-Legacy, GPT-3.5-Turbo, and GPT-4	Engineering fundamental knowledge	An unpublished practice exam	Closed-ended (MSQs, MMQs, and filling in the blank)	134

(*continued*)

Table 2. Continued

No	References	GPLLMs	Specific domain	Test datasets: Question sources	Question types	Number of questions
16	Rosól *et al.* (2023)	GPT-3.5 and GPT-4	Medicine	3 versions of PMFE	Closed-ended (MSQs)	600
17	Saad *et al.* (2023)	GPT-4	Orthopedics	Mock FRCS Orth Part A	Closed-ended (MSQs)	240

Note(s): CogTale: Cognitive Treatments Article Library and Evaluation FEBU: Fellow of the European Board of Urology; FRCS Orth: Orthopedic fellow of the Royal College of Surgeons; iDISK: International Dietary Supplement Knowledgebase; KGSBE: Korean General Surgery Board Exams; MSKCC: Memorial Sloan Kettering Cancer Center; MSRA: Multi-Specialty Recruitment Assessment; NBME: National Board of Medical Examiners; OITE: Orthopedic In-Training Examination; PMFE: Polish Medical Final Examination; TNSPBE: Turkish Neurosurgical Society Proficiency Board Exams
Source(s): Authors' own work

authority, answer judgment, and benchmark suitability. Therefore, closed-ended questions are commonly used in research, while open-ended questions are less frequently employed in evaluating the performance of GPLLMs (Table 2). Most existing studies utilize hundreds of closed-ended questions for QA test dataset development, including multiple-choice questions (Sahin *et al.*, 2024; Hou and Zhang, 2024; Rizzo *et al.*, 2024; Schoch *et al.*, 2024; Tsoutsanis and Tsoutsanis, 2024; Antaki *et al.*, 2023), true/false judgments (Hou and Zhang, 2024; Rasool *et al.*, 2024), and fill-in-the-blank questions (Mansurova *et al.*, 2024; Pursnani *et al.*, 2023). Although numerous test datasets exist, there remains a lack of standardized benchmark datasets specifically designed to evaluate and compare CLQA performance.

2.3 Research gaps

Existing studies (Table 1) on construction-related QA have laid the foundation for this work. However, they exploit conventional deep learning or machine learning QA models, which have disadvantages concerning (1) the requirement of large amounts of training data and (2) a small-scale QA scope. In contrast, CLQA involves a broad spectrum of subareas (e.g. permits, contracting, safety, disputes, quality, and environment), and it also faces challenges in acquiring sufficient QA pairs for training learning-based models. Hence, existing studies using conventional learning-based models are inapplicable to achieving CLQA.

Although GPLLMs exhibit higher language processing capacity than conventional learning-based models and hold the advantage of not requiring training, their lack of domain-specific knowledge may compromise CLQA performance (Table 2). Developing domain knowledge bases faces three challenges. First, the development process heavily depends on an expert-driven selection of documents containing domain knowledge, which limits the scope and updates of CLKR. Second, there is insufficient empirical evidence to confirm that domain knowledge improves CLQA performance. Third, there is a lack of benchmark datasets to test and compare CLQA performance.

3. Methodology

To resolve the aforementioned challenges, a 4-phase methodology is devised to build a CLKR to lift the CLQA performance of GPLLMs. The first phase involves obtaining 702 candidate documents from a total of 374,992 written judgments. In the second phase, 387 CL documents are filtered to ensure coverage across eight distinct CL areas and form the CLKR. The third phase focuses on integrating the CLKR with GPLLMs using the LangChain-based RAG technology. The final phase compares the performance of seven GPLLMs, both with and

without the CLKR, by utilizing a 2,140-question CLQA validation set. This comparison not only measures the GPLLMs' performance improvements but also validates the effectiveness of the CLKR (Figure 1).

3.1 Recognition of candidate documents for CLKR

This phase (Figure 1) is designed to collect candidate documents for constructing the CLKR in a data-driven manner. It encompasses (1) gathering corpora (i.e. written judgments) containing construction laws, (2) identifying CL document name entities within the written judgments, and (3) cleansing these CL document name entities. Further details are provided in Figure 2a.

The collection of construction-related legal corpora consists of determining corpora sources and clarifying search keywords. The written judgments of the construction industry encompass various laws referred to in real-world legal cases, which could be an ideal and authoritative source for determining candidate documents. For example, the "China Judgments Online" is established and developed by the government to show all legal case judgments in Mainland China to the public, and its authority and comprehensiveness ensure both the quality and quantity necessary for serving as the corpora source (SPC, 2023). This study uses the keywords "construction engineering" and "construction project" to search for written judgments in the construction industry. These keywords can appear in any part of the judgment text. Considering the extensive historical corpora, this study focuses on the most recent legal cases. As a result, a total of 374,992 written judgments from January 2021 to July 2023 are collected through online retrieval (Figure 2a).

Subsequently, document name entities are recognized in 374,992 written judgments (Figure 2a). Guillemets are a pair of punctuation marks in the form of sideways double chevrons (i.e. 《》), which are used to highlight names of books, articles, laws, regulations, and other documents in Chinese text. This study employs guillemets as identifiers to precisely and efficiently recognize document entities (e.g. 《Management Regulations of Registered Construction Engineers》). Utilizing this identifier-based recognition method, a total of 772,559 document name entities are retrieved from these judgments (Table S1).

The cleansing process of document name entities involves merging identical entities, excluding low-frequency items, and eliminating non-law documents (Figure 2a). Specifically,

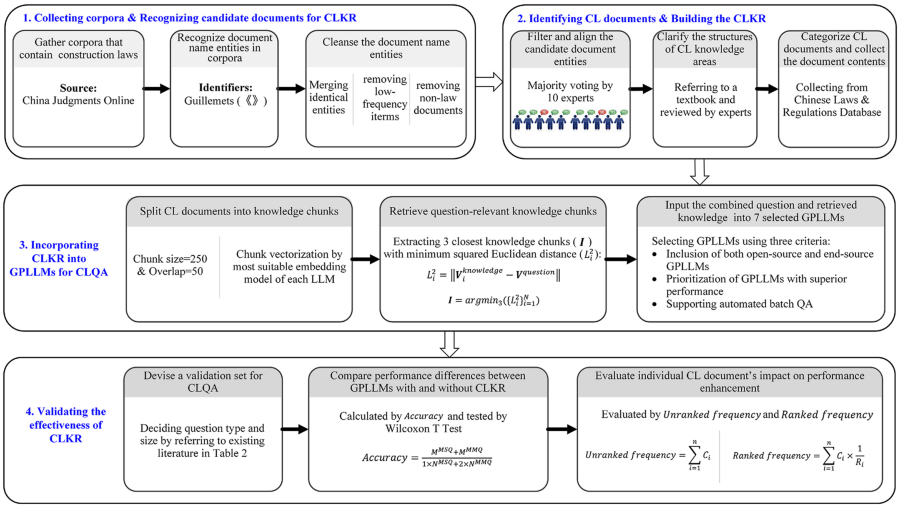


Figure 1. The phases of building a CLKR to lift the CLQA performance of GPLLMs. **Source(s):** Authors' own work

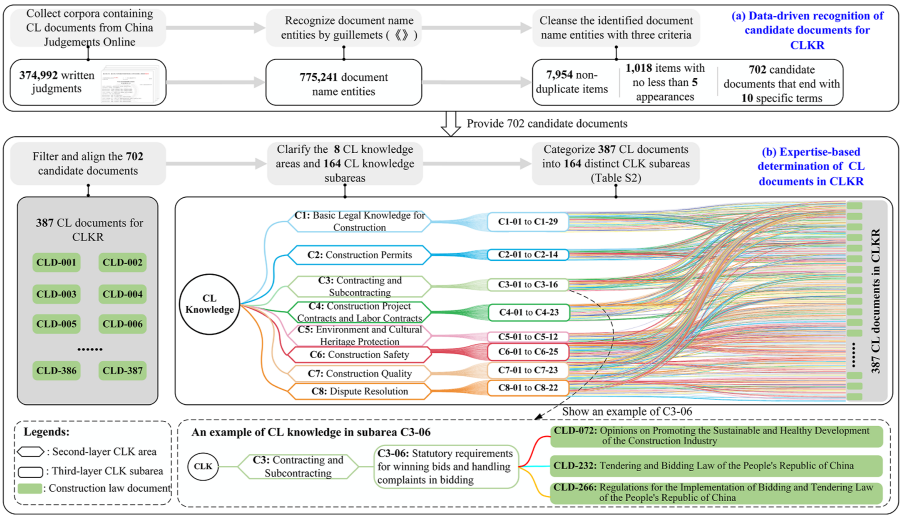


Figure 2. Building the CLKR by combining data-driven and expertise-based paradigm. **Source(s):** Authors' own work

duplicates among 772,559 document name entities are merged, and 7,954 non-duplicate items are retained (Figure 2a). Then, the document name entities that appear less than five times (i.e. an occurrence rate of less than 0.01% in 374,992 cases) are excluded, significantly reducing the name entities containing various typographical errors. This reduces the number of documents from 7,954 to 1,018 (Figure 2a). As the naming system for legal documents in Mainland China is rigorous, legal document names must end with 10 specific terms, including decision/ruling, guidelines, interpretation, law/code, method/procedure, notice, opinion/advisory, ordinance, regulation/rule, and standards/norm. Thus, non-legal documents lacking these keywords are excluded, such as ⟪Shenzhen Newspaper⟫ and ⟪Commercial Housing Sales Contract⟫. After these exclusions, 702 documents remain as candidates for CLKR construction (Figure 2a). Table S1 shows the whole process from 374,992 written judgments to 702 candidate documents step by step.

3.2 Determination of construction laws in CLKR

Following the acquisition of 702 candidate documents through the data-driven approach, expertise is applied in this phase to filter the CL documents and construct the CLKR (Figure 2b). The procedure includes (1) filtering out and aligning the document entities, (2) refining a "1-8-164" structure of CL knowledge areas, and (3) categorizing the CL documents and gathering the specific contents.

This step involves filtering out ambiguous named entities and aligning different entities that refer to the same CL document (Figure 2b). Ten experienced experts (Table S1) are invited to manually review 702 documents one by one. A majority vote is conducted for each removal to minimize the subjectivity of the experts' judgments. If 6 or more out of 10 experts agree that a document should be excluded, it is removed from the list of candidate documents. Ambiguous named entities do not specify a specific document, such as "notice", "guidelines", and "interpretation", which account for 17 out of the 702 document name entities excluded (Table S1). Subsequently, multiple entities referring to the same CL document are aligned, and redundant document entities are removed. For example, "Civil Code" is removed as it refers to the same entity as "Civil Code of the People's Republic of China". As a result, 298 of the 685

candidate documents are excluded, leaving 387 CL documents in the final set (Figure 2b). The voting details and the documents excluded and retained are both attached in Table S1.

A three-layer knowledge hierarchy (Figure 2b) has been established by referring to the textbooks (Li *et al.*, 2021) and a group of ten experts (Table S1). The first layer is the root node entitled CL knowledge, beneath which lie 8 second-layer knowledge areas concerning basic legal knowledge, permitting, contracting and subcontracting systems, project and labor contracts, protection, safety, quality, and disputes (Figure 2b). These 8 second-layer areas are further subdivided into 164 third-layer CL knowledge subareas to organize the documents in CLKR with a finer granularity (Figure 2b and Table S2).

After establishing the three-layer knowledge hierarchy, the experts categorized 387 CL documents into 164 distinct third-layer CL knowledge subareas (Figure 2b and Table S2). For instance, the CL knowledge subarea C3-06 has three CL documents (CLD): CLD-072, CLD-232, and CLD-266. Each CL document can be categorized into multiple knowledge subareas. This categorization process clarifies the relationships between CL knowledge subareas and CL documents, which further ensures that CLKR comprehensively covers all 164 CL knowledge subareas. The specific contents of these CL documents are obtained from the Chinese Laws & Regulations Database (NPC, 2024). Finally, the CLKR, containing 387 documents across 8 second-layer areas and 164 third-layer subareas, has been developed and released in the GitHub repository.

3.3 CLKR-empowered GPLLMs for CLQA

This section utilizes the RAG to facilitate integration between the CLKR and GPLLMs for CLQA. It comprises three steps (Figure 1), including (1) dividing the 387 CL documents from the CLKR into knowledge chunks (Figure 3a), (2) retrieving knowledge chunks pertinent to the CL question (Figure 3b), and (3) integrating the question with the retrieved knowledge for input into seven selected GPLLMs (Figure 3c).

RAG is a widely adopted method for integrating external knowledge into GPLLMs (Mansurova *et al.*, 2024; Alan *et al.*, 2024). This study employs the extensively used LangChain framework for the RAG (Langchain-ai, 2024). The CLKR is first loaded using a document loader, and the loaded documents are then divided into smaller knowledge chunks (Figure 3a), instead of presenting the entire document (Rasool *et al.*, 2024; Alan *et al.*, 2024). Chunk size refers to the token count of each chunk during the document dividing process, while overlap indicates the repeated tokens between adjacent chunks (Mansurova *et al.*, 2024; Alan *et al.*, 2024). After referring to existing studies (Langchain-ai, 2024; Eleliemy and Ciorba, 2021; Wang *et al.*, 2024), the chunk size is set to 250 tokens. Additionally, certain GPLLMs (e.g. Llama-2-70B) have a 4096-token limit for one input (meta-llama, 2024), which includes three knowledge chunks, the prompt, and the question itself. Setting each chunk size to 250 words ensures that the total length of these elements stays within this limit (meta-llama, 2024; Wang *et al.*, 2024). The adjacent chunks are designed to overlap by 50 tokens to maintain data continuity and prevent information loss at the boundaries of the chunks (Langchain-ai, 2024; Domengo, 2024; Eleliemy and Ciorba, 2021). The CL knowledge chunks are then vectorized using embedding models, changing from text form into numerical knowledge vectors (Figure 3a). In selecting embedding models, the author adheres to the recommendations from the GPLLM developer, as these recommended embedding models allow GPLLM to attain peak performance (Rasool *et al.*, 2024; Hou and Zhang, 2024). For example, the embedding model of ERNIE Embedding-V1 is recommended for ERNIE-Bot 4.0 in its official technical documentation (Table 3). *These CL knowledge vectors are stored in a FAISS-formatted vector store for reuse. Both FAISS and Chroma are the most commonly used vector stores, but FAISS is particularly noted for its faster retrieval speed in RAG tasks* (Langchain-ai, 2024; Rasool *et al.*, 2024).

Then, this study exploits Euclidean distance-based method to find question-related knowledge chunks from the *FAISS* vector store (Figure 3b). It is defined as:

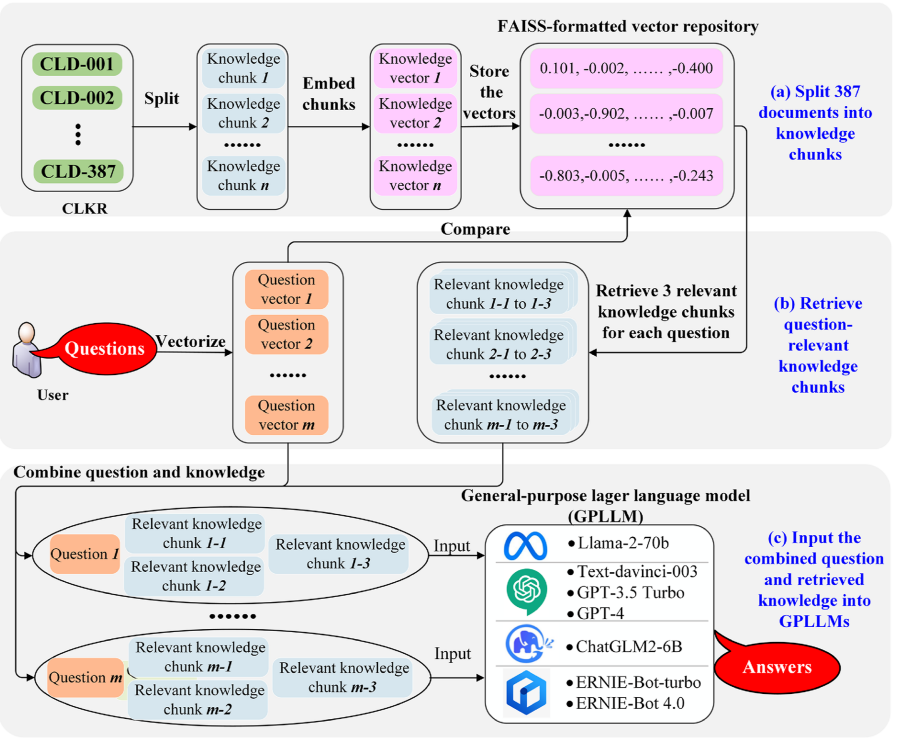


Figure 3. The process of leveraging CLKR to empower GPLLM for CLQA. **Source(s):** Authors' own work

Table 3. GPLLMs selected for integration with CLKR

No	Contributors	GPLLMs	Parameters	Best in processing	Open-/ Closed-source	Corresponding embedding models
1	Meta	Llama-2-70b	70 billion	English	Open-source	all-mpnet-base-v2
2	OpenAI	text-davinci-003	Unknown	English	Closed-source	text-embedding-ada-002
3		GPT-3.5 Turbo	20 billion	English	Closed-source	text-embedding-ada-002
4		GPT-4	1.8 trillion	English	Closed-source	text-embedding-ada-002
5	Tsinghua University	ChatGLM2-6B	6 billion	Chinese	Open-source	text2vec-large-chinese
6	Baidu	ERNIE-Bot-turbo	13 billion	Chinese	Closed-source	ERNIE Embedding-V1
7		ERNIE-Bot 4.0	>1 trillion	Chinese	Closed-source	ERNIE Embedding-V1

Source(s): Authors' own work

$$L_i^2 = \left\| V_i^{knowledge} - V^{question} \right\| \tag{1}$$

where L_i^2 represents the squared Euclidean distance between the ith knowledge chunk $V_i^{knowledge}$ and the question vector $V^{question}$. A smaller distance signifies a higher similarity, indicating that

the knowledge chunk is more relevant to the question. The three knowledge vectors nearest to the question vector are defined mathematically as:

$$\boldsymbol{I} = argmin_3\left(\left\{L_i^2\right\}_{i=1}^{N}\right) \tag{2}$$

where I is the set of indices for the three closest knowledge vectors. The selected knowledge chunks are subsequently utilized as the background information for addressing the query.

Finally, the top three relevant knowledge chunks and the question are combined as a new query, and the query is then inputted into GPLLM to get the corresponding answer (Figure 3c). In this study, seven GPLLMs are selected for test. The reasons for selecting these GPLLMs are: (1) inclusion of both open-source and closed-source GPLLMs, (2) prioritization of GPLLMs demonstrating superior performance, and (3) a requirement that the GPLLMs support automated batch QA. The chosen GPLLMs in Table 3 represent a mix of open-source and closed-source technologies. As indicated on the LLM leaderboard, OpenAI and Baidu are recognized as the leading closed-source models for English and Chinese respectively (Pei *et al.*, 2024; Oh *et al.*, 2023). Similarly, Meta and Zhipu stand out as the top open-source models for English and Chinese (Pei *et al.*, 2024; Lu *et al.*, 2024). Since this study requires each model to answer thousands of questions, only GPLLMs that are either open-source or closed-source models with accessible APIs can be included (Sahin *et al.*, 2024; Saad *et al.*, 2023; Gilson *et al.*, 2023). As a result, certain popular GPLLMs like Copilot, which fundamentally rely on OpenAI's GPLLMs and do not support automated batch QA, have to be excluded from this study (Khan, 2024).

3.4 Effectiveness validation of CLKR

The validation of CLKR's effectiveness is determined by examining whether GPLLMs have performance enhancements before and after the integration with CLKR (Figure 1). It involves (1) devising a validation set for CLQA, (2) comparing the differences between initial GPLLMs and CLKR-empowered GPLLMs, and (3) evaluating individual CL document impact on CLQA performance enhancement.

As there are no ready-to-use benchmark datasets of CLQA, a validation set comprising 2,140 real questions is developed, which covers first-level PCEQEs (11 test papers) and second-level PCEQEs (13 test papers) from 2014 to 2023 (Figure 4). The PCEQE is the most authoritative qualification assessment for those aspiring to be registered construction engineers in Mainland China (Liu and Low, 2011). Of the 2,140 questions, 1,550 are multiple-choice single-answer questions (MSQs), and 590 are multiple-choice multiple-answer (MMQs) (Figure 4). Additionally, each question is labeled with the sourced PCEQE paper and the corresponding CL knowledge area (Table S3).

The performance differences between GPLLMs with and without are calculated by accuracy and statistically tested using the Wilcoxon T Test. Since MSQ and MMQ vary in difficulty, they are assigned different point values in PCEQEs. Each MSQ is worth one point, and each MMQ is worth two points. For MMQs, full marks of two points are awarded if all correct options are selected. Choosing some, but not all, correct options can earn 0.5 points for each correct option chosen. Selecting any incorrect option results in zero points. The CLQA accuracy of the GPLLM is assessed based on the marks it receives on PCEQE papers or particular question sets (e.g. all 457 questions in the C1 subarea as shown in Figure 4), which is specifically defined as:

$$Accuracy = \frac{M^{MSQ} + M^{MMQ}}{1 \times N^{MSQ} + 2 \times N^{MMQ}} \tag{3}$$

where M^{MSQ} and M^{MMQ} refer to the marks obtained on MSQs and MMQs, and N^{MSQ} and N^{MMQ} are the number of MSQs and MMQs.

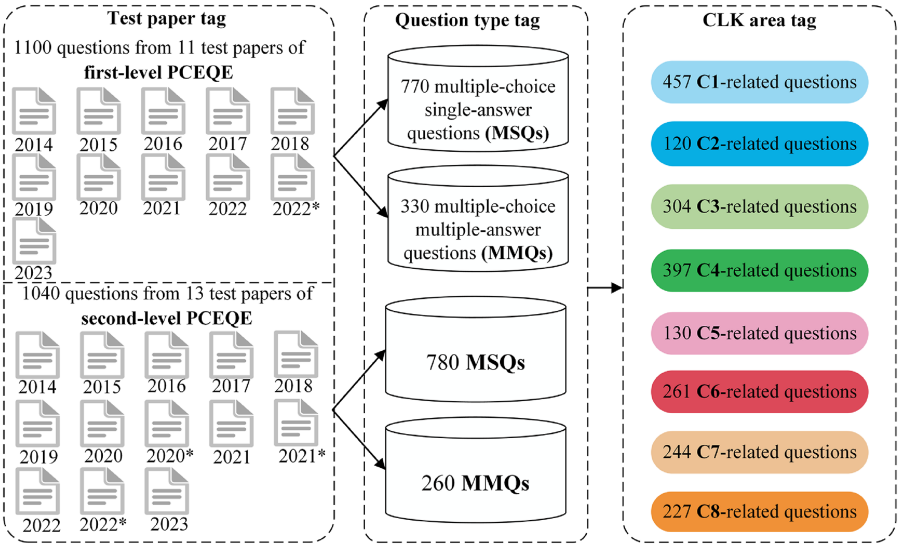


Figure 4. The CLQA dataset. Note: * indicates an extra PCEQE hold that year. **Source(s):** Authors' own work

Additionally, the Wilcoxon *T* Test is used to test whether there is a significant difference between the performance obtained by initial GPLLMs and CLKR-empowered GPLLMs (Figure 1). The mark comparisons are conducted from three perspectives: 24 PCEQE papers, MSQs/MMQs, and 8 second-layer knowledge areas. If CLKR-empowered GPLLMs show significant performance improvements compared to the original GPLLMs in PCEQEs and across CL knowledge areas, this could strongly validate the effectiveness and comprehensiveness of CLKR. Conversely, if there is no significant difference, then CLKR is ineffective.

Finally, the impact of each of the 387 CL documents on CLQA performance improvement is quantitatively evaluated using two distinct indicators: unranked frequency (Eq. (4)) and ranked frequency (Eq. (5)). For each of the 2,140 questions, three question-related knowledge chunks from the 387 documents will be extracted. Unranked frequency measures how often a document is referenced, counting every appearance of the knowledge chunk-sourced documents regardless of the similarity rank. It is defined as:

$$Unranked\,frequency = \sum_{i=1}^{n} C_i \tag{4}$$

where $C_i = 1$ if a chunk-sourced document appears, otherwise $C_i = 0$; n refers to the total number of retrieved knowledge chunks. Ranked frequency considers the rank of similarity when counting knowledge chunk-sourced document appearances, which is defined as:

$$Ranked\,frequency = \sum_{i=1}^{n} C_i \times \frac{1}{R_i} \tag{5}$$

where R_i refers to the rank (i.e. 1, 2, or 3) of the top 3 question-relevant knowledge chunks.

4. Results

With the devised four-phase methodology, this study obtains the answers from seven pairs of original and CLKR-empowered GPLLMs, as well as the extracted knowledge chunks during CLKR-engaged CLQA, as depicted in Figure 5. All 29,960 answers provided by these GPLLMs and the knowledge chunks with their similarity ranks are found in Table S4. The authors compare CLQA accuracies of the original versus CLKR-enhanced GPLLMs on different test papers, between MSQs and MMQs, across 8 CL knowledge areas, and in 100 open-ended questions (Figure 5). Additionally, each CL document's impact is evaluated and compared by unranked and ranked frequency (Figure 5).

4.1 Performance comparison between GPLLMs with and without CLKR

4.1.1 CLKR-enabled performance enhancements on PCEQE test papers. The Wilcoxon *T* Test results demonstrate that CLKR significantly enhances the accuracy of seven different GPLLMs in CLQA (Table 4). On average, the CLKR results in a remarkable 21.1% increase in the accuracy of these GPLLMs (Table 4). The performance enhancement across the seven GPLLMs ranges from 9.9% to 44.9% (Figure 6). The CLKR-empowered text-davinci-003 exhibits the most substantial improvement, achieving a 44.9% accuracy increase, rising from 0.329 to 0.476 (Figure 6b). Despite the CLKR-empowered ERNIE-Bot 4.0 shows the least improvement at only 9.9% (Figure 6g), the effectiveness of CLKR is also significantly confirmed by Wilcoxon *T* Test (Table 4).

CLKR can significantly enhance the CLQA performance of GPLLMs regardless of their NLP capabilities and their training language (i.e. Chinese or English) (Tables 3 and 4). ERNIE-Bot 4.0 leads the performance of individual GPLLMs, with an improvement in accuracy from

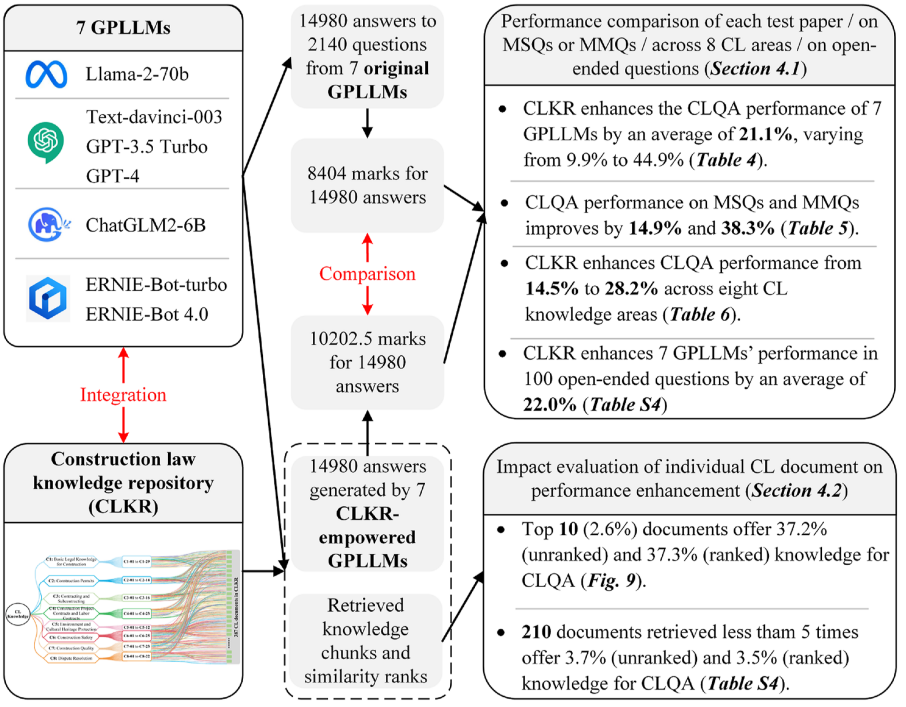


Figure 5. The comparison results of GPLLMs' CLQA accuracies with and without CLKR. **Source(s):** Authors' own work

Table 4. Wilcoxon *T* Tests on CLQA accuracy of 7 GPLLMs with and without CLKR in PCEQEs

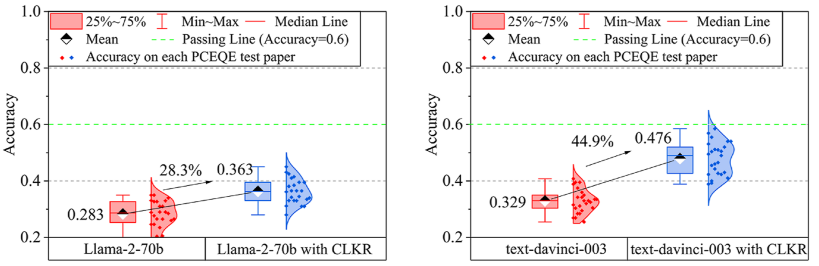
No	GPLLM	CLKR	Average accuracy	Accuracy enhancement	*z*-statistic	*p*-value
1	Llama-2-70b	without	0.283	28.3%	4.197	0.000***
		with	0.363			
2	text-davinci-003	without	0.329	44.9%	4.286	0.000***
		with	0.476			
3	GPT-3.5 Turbo	without	0.349	36.3%	4.287	0.000***
		with	0.476			
4	GPT-4	without	0.528	25.4%	4.171	0.000***
		with	0.663			
5	ChatGLM2-6B	without	0.430	11.1%	3.729	0.000***
		with	0.478			
6	ERNIE-Bot-turbo	without	0.419	10.2%	3.429	0.002***
		with	0.462			
7	ERNIE-Bot 4.0	without	0.755	9.9%	4.029	0.000***
		with	0.830			
Average accuracy of 7 GPLLMs		without	0.442	21.1%	NA	NA
		with	0.535			

Note(s): *** denote confidence levels above 99%
Source(s): Authors' own work

0.755 to 0.830 (Figure 6g). The CLKR-empowered GPT-4 also shows significant gains, with its accuracy increasing from 0.528 to 0.663 (Figure 6d), surpassing the PCEQE passing mark of 0.6. Although the average accuracy of other 5 GPLLMs does not pass the PCEQE tests, their considerable improvements also affirm the effectiveness of CLKR in boosting the performance of GPLLMs in CLQA (Figure 6a, 6b, and 6c and 6e-6f). Meanwhile, the CLKR built based on CL documents written in Chinese (Table S2) significantly boosts the CLQA performance of GPLLMs launched by Chinese institutions (i.e. ChatGLM2-6B, ERNIE-Bot-turbo, and ERNIE-Bot 4.0) (Figure 6e, 6f, and 6g). It also provides notable improvements for GPLLMs primarily trained in English corpora like Llama-2-70b, text-davinci-003, GPT-3.5 Turbo, and GPT-4 (Figure 6a, 6b, 6c, and 6d).

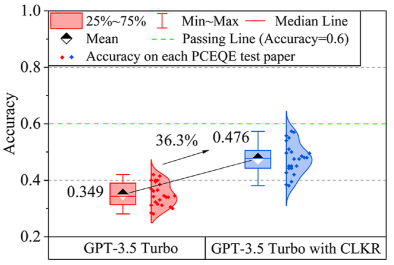
4.1.2 CLKR-enabled performance enhancements in MSQs and MMQs. In the CLQA performance comparative analysis (Table 5), the integration of CLKR significantly enhances the performance of 7 GPLLMs in answering both types of multiple-choice questions (i.e. MSQs and MMQs) (Figure 7). Specifically, the accuracy of GPLLMs on MSQs improves by 14.9%, increasing from 0.569 to 0.654 (Table 5). Text-davinci-003 demonstrates the most significant enhancement in MSQs among all GPLLMs, achieving an improvement of 40.4% (Figure 7b). Meanwhile, GPT-3.5 Turbo exhibits the highest improvement in MMQs (Figure 7c), with an impressive increase of 86.2%. Although the accuracy of GPLLMs on MSQs is higher than on MMQs (Figure 7), the improvement ratio for MMQs is greater at 38.3%, increasing from 0.273 to 0.378 (Table 5). The discussions about the performance difference between MSQs and MMQs are conducted in Section 5.3. Among the GPLLMs evaluated (Table 5), the CLKR-empowered ERNIE-Bot 4.0 stands out as the top performer, showing superior capability in handling both MSQs and MMQs, with its accuracy on MSQs even exceeding 0.9 (Figure 7g). However, it recorded the lowest improvement among the seven GPLLMs in MSQs, achieving an increase of only 6.6% (Figure 7g).

4.1.3 CLKR-enabled performance enhancements across 8 CL knowledge areas. The results reveal that the CLKR significantly lifts the CLQA accuracy of GPLLMs across eight CL knowledge areas (C1-C8). Figure 8 visually shows the extent to which CLKR-empowered GPLLMs have improved their question-answering capabilities in each area, despite varying degrees of improvements (Figure 8a, 8b, 8c, 8d, 8e, 8f, and 8g). For example, GPT-4.0's

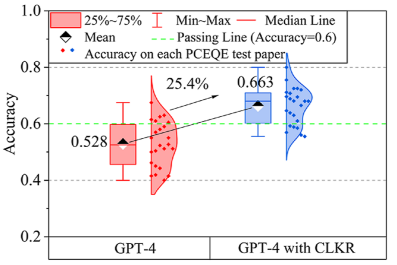


(a) Accuracy of Llama-2-70b with and without CLKR

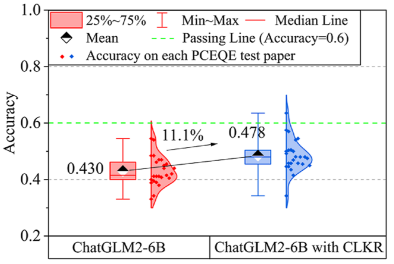
(b) Accuracy of text-davinci-003 with and without CLKR



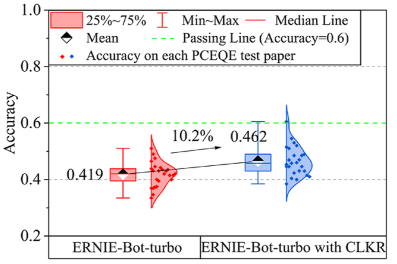
(c) Accuracy of GPT-3.5 Turbo with and without CLKR



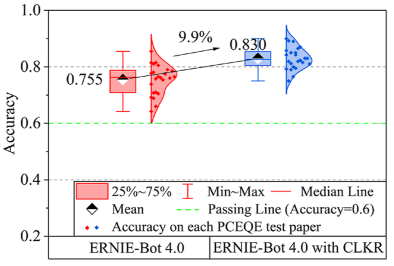
(d) Accuracy of GPT-4 with and without CLKR



(e) Accuracy of ChatGLM2-6B with and without CLKR



(f) Accuracy of ERNIE-Bot-turbo with and without CLKR



(g) Accuracy of ERNIE-Bot 4.0 with and without CLKR

Figure 6. Performance of original and CLKR-empowered GPLLMs in PCEQEs. **Source(s):** Authors' own work

accuracy enhancement is the minimum in C1: basic legal knowledge for construction (14.7%) and the largest in C2: construction permits (41.7%) (Figure 8d). Each GPLLM shows different maximum or minimum CL knowledge area-specific accuracy improvement (Figure 8a, 8b, 8c,

Table 5. Wilcoxon *T* Tests on CLQA accuracy of GPLLMs with and without CLKR in MSQs and MMQs

Question type	CLKR	Average accuracy	Accuracy enhancement	*z*-statistic	*p*-value
MSQs	without	0.569	14.9%	9.451	0.000***
	with	0.654			
MMQs	without	0.273	38.3%	9.360	0.000***
	with	0.378			

Source(s): Authors' own work

8d, 8e, 8f, and 8g). The CL knowledge area-specific accuracy improvements range from 14.5% to 28.2% (Table 6). These notable increases in CLQA accuracy (Table 6 and Figure 8) not only validate the CLKR's efficacy but also underscore CLKR's comprehensive coverage of CL knowledge areas.

Across 8 CL knowledge areas, ERNIE-Bot 4.0 demonstrates the highest CLQA accuracy among all individual GPLLMs (Figure 8a, 8b, 8c, 8d, 8e, 8f, and 8g), achieving 5.6%–15.8% improvements in each area with the integration of CLKR (Figure 8g). GPT-4 shows accuracy below 0.6 across all CL knowledge areas before CLKR incorporation (Figure 8d), whereas the accuracy enhancement brought by CLKR enables it to exceed the PCEQE passing threshold of 0.6 in 8 CL knowledge areas (Figure 8d). Other 5 GPLLMs do not reach the 0.6 passing threshold across each area, however, the integration of CLKR brings marked accuracy improvements across all CL knowledge areas (Figure 8a, 8b, and 8c and 8e-8f).

4.1.4 CLKR-enabled performance enhancements on open-ended questions. CLKR not only improves GPLLMs' performance in closed-ended questions but also enables an average accuracy improvement of 22.0% across seven GPLLMs in open-ended questions (Table S4). Figure 9 visually presents the accuracies of 6 pairs of GPLLMs on 100 open-ended questions, which are converted from 100 multiple-choice questions in the test dataset by removing the choices (Table S4). Three experienced experts are invited to evaluate the responses of GPLLMs to open-ended questions, scoring each response either 0 or 1, with the final scores being the average of the three experts' ratings (Table S4). Although no GPLLM achieves an accuracy exceeding 0.6 (the passing threshold of PCQEQs) even with the integration of CLKR, the performance enhancements enabled by CLKR remain considerable (Figure 9).

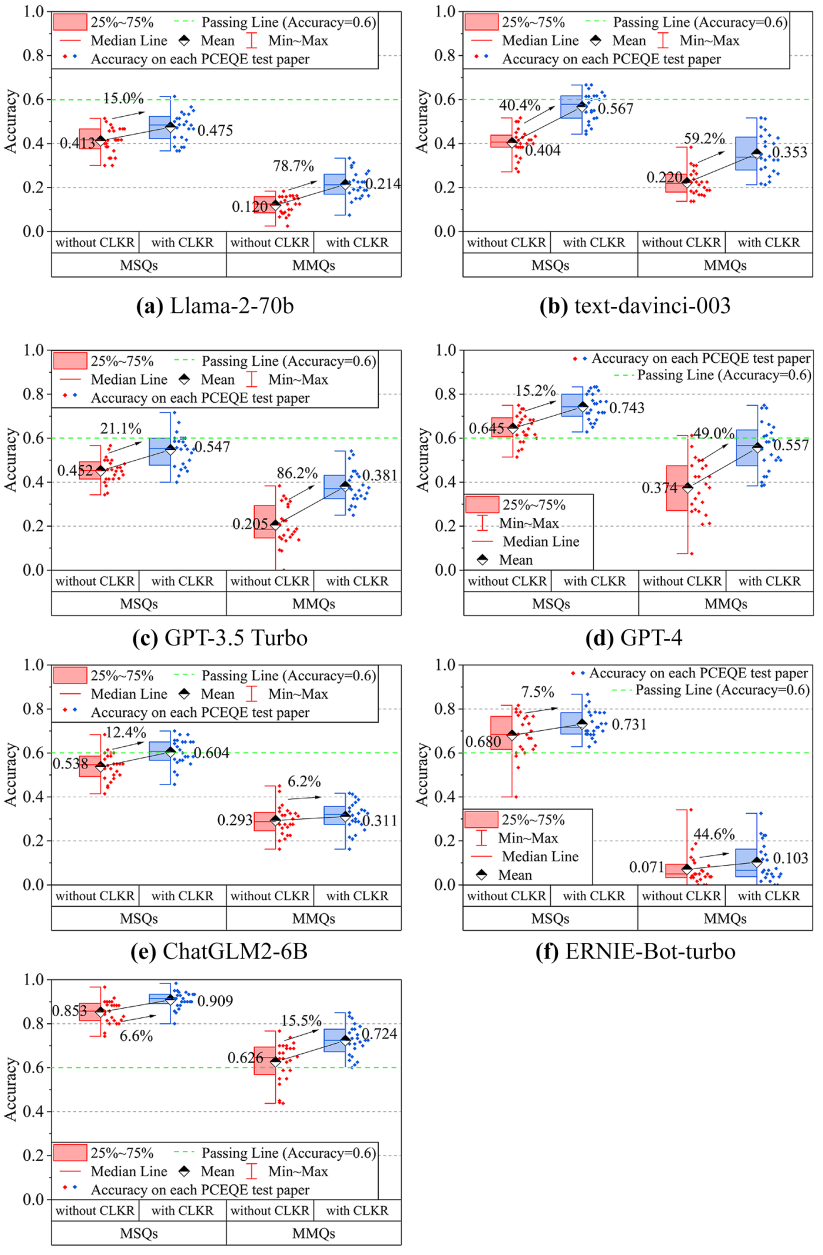
4.2 The impact evaluation of each CL document on performance enhancement

A long-tail effect among CL documents in the CLKR has been observed, which conforms to a power law distribution (Figure 10). The top 10 of 387 (2.6%) documents offer 37.2% (Figure 10a) and 37.3% (Figure 10b) contextual knowledge for CLQA under unranked frequency and ranked frequency. The three most frequently retrieved documents are CLD-260: Regulations on the Management of Construction Project Quality, CLD-380: Civil Code of the People's Republic of China, and CLD-347: Regulations on the Causes of Action for Civil Cases (Figure 10a). Concurrently, 210 CL documents are retrieved fewer than five times, and 101 CL documents do not even make any contribution to the CLQA (Table S4). Nonetheless, despite the minimal individual contributions from most documents, their collective significance as the long tail (Figure 10) remains crucial for enhancing GPLLMs' CLQA performance.

5. Discussion

5.1 Mis-answered question types and potential reasons for imperfect CLQA performance

While incorporating CLKR into GPLLMs has effectively improved their CLQA performance, the accuracy remains suboptimal (Figure 6–9). After analyzing the incorrect answers, it is



(a) Llama-2-70b **(b)** text-davinci-003

(c) GPT-3.5 Turbo **(d)** GPT-4

(e) ChatGLM2-6B **(f)** ERNIE-Bot-turbo

(g) ERNIE-Bot 4.0

Figure 7. Performance comparison of original and CLKR-empowered GPLLMs in MSQs and MMQs. **Source (s):** Authors' own work

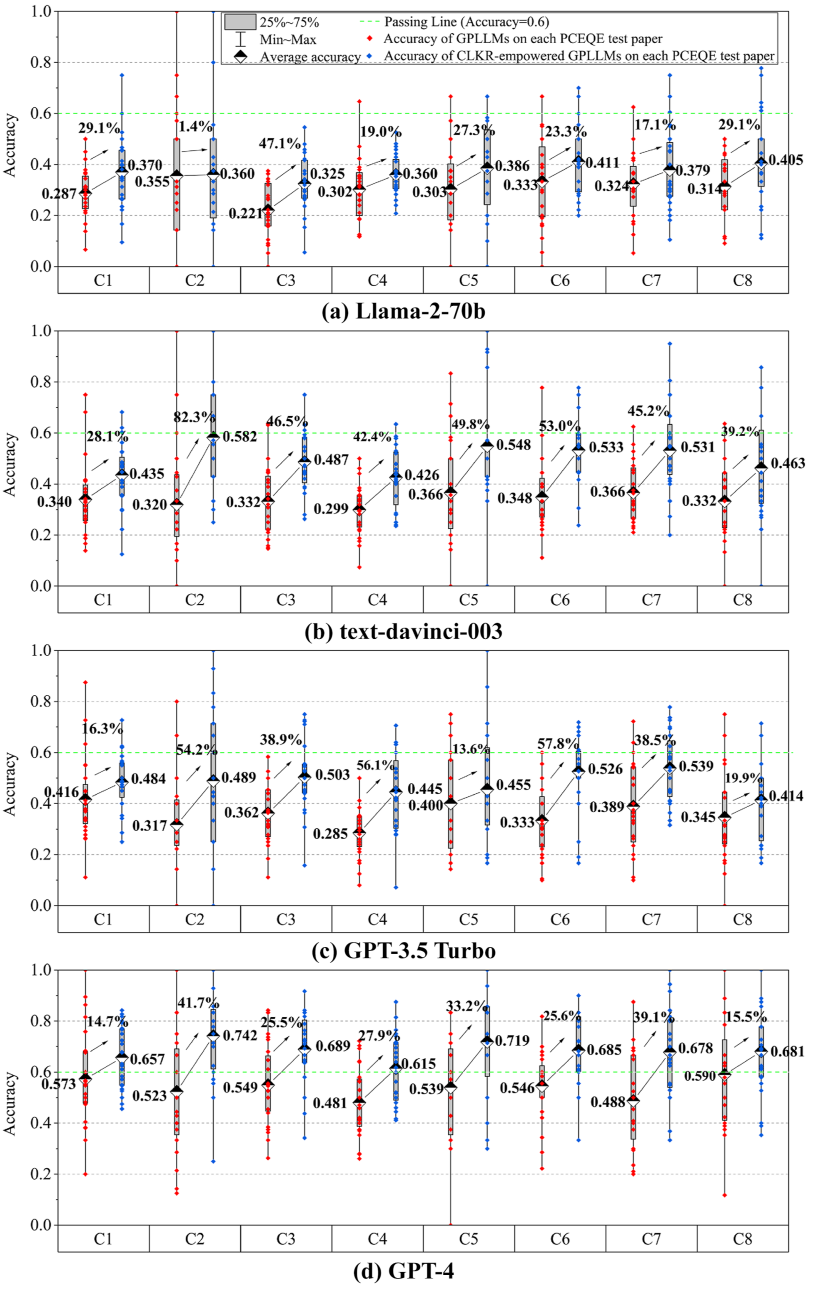


Figure 8. Performance comparison of original and CLKR-empowered GPLLMs across C1-C8. **Source(s):** Authors' own work

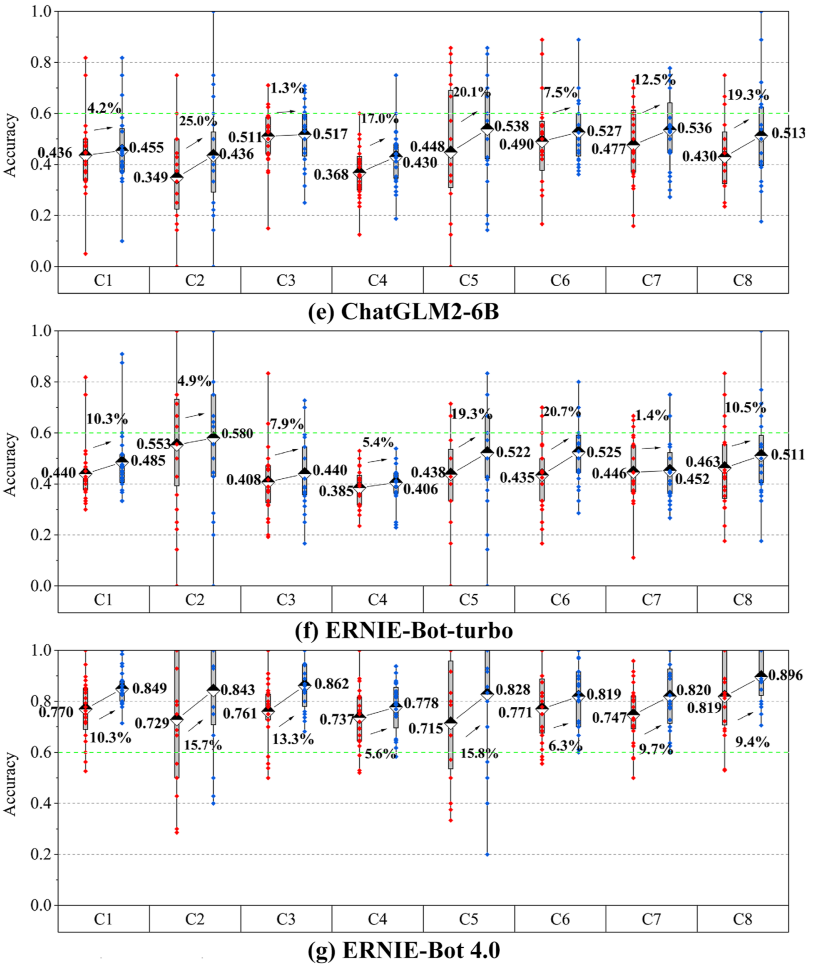


Figure 8. (*continued*)

found that errors are more likely to occur in multiple-answer, scenario-based, and long-text questions. Beyond the inherent complexity of MMQs compared to MSQs, the main reasons for imperfect performance are (1) insufficient scenario analysis ability and (2) poor understanding of long-text contexts.

The insufficient scenario analysis ability refers to the notably lower accuracy in answering scenario-based questions (SBQs) compared to recall-based questions (RBQs) (Table S4). SBQs typically do not explicitly specify the knowledge points required for the answer but rather present a hypothetical scenario (Saka *et al.*, 2024; Badyal *et al.*, 2023), as opposed to simply recalling specific facts which are defined as RBQs (Badyal *et al.*, 2023; Hwang and Mattila, 2022). The integration of CLKR improves GPLLMs' accuracy by 23.0% on RBQs but only 4.6% on SBQs, with the latter's minimal enhancement limiting overall CLQA performance (Table S4). Future efforts to create specific libraries of scenarios within the CL

Table 6. Wilcoxon *T* Tests on CLQA accuracy of GPLLMs with and without CLKR across C1-C8

Knowledge domain	CLKR	Average accuracy	Accuracy enhancement	*z*-statistic	*p*-value
C1	without	0.466	14.5%	6.672	0.000***
	with	0.534			
C2	without	0.449	28.2%	5.896	0.000***
	with	0.576			
C3	without	0.449	21.6%	6.825	0.000***
	with	0.546			
C4	without	0.408	21.1%	7.086	0.000***
	with	0.494			
C5	without	0.458	24.5%	5.966	0.000***
	with	0.571			
C6	without	0.465	23.6%	7.427	0.000***
	with	0.575			
C7	without	0.462	21.6%	7.334	0.000***
	with	0.562			
C8	without	0.470	17.9%	5.970	0.000***
	with	0.555			

Source(s): Authors' own work

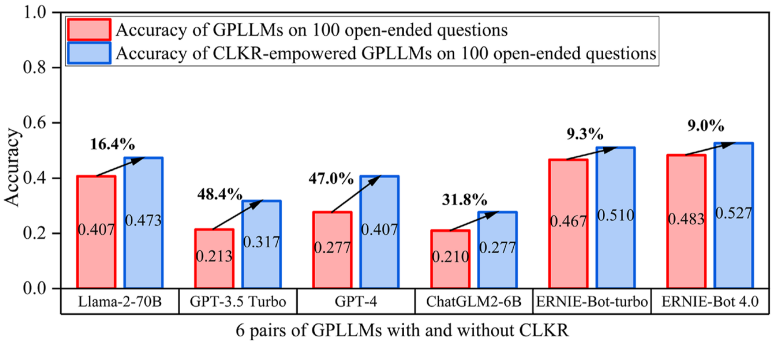
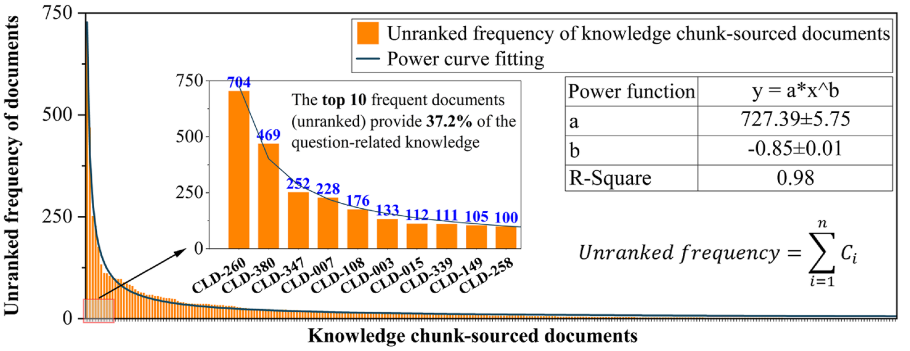


Figure 9. Performance comparison of GPLLMs with and without CLKR in 100 open-ended questions. Note: The API of text-davinci-003 model is no longer accessible, when the authors conduct the CLQA on the open-ended question set in Nov. 2024. **Source(s):** Authors' own work

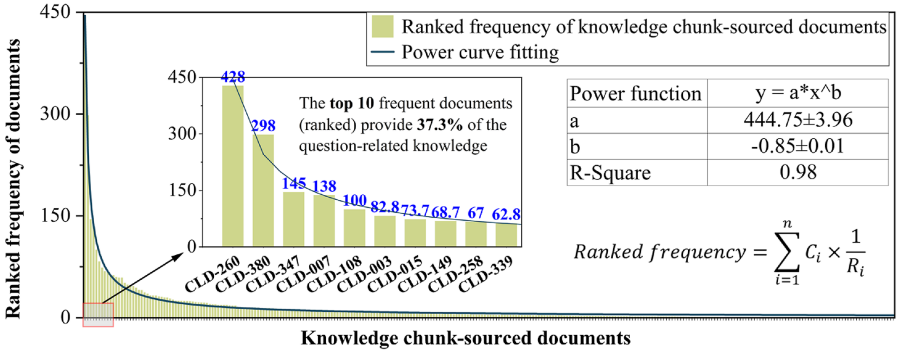
domain may offer a path toward a substantial enhancement of GPLLMs' scenario analysis capabilities to achieve optimal CLQA performance (Saka *et al.*, 2024; Mansurova *et al.*, 2024).

Pearson correlation tests reveal a strong negative correlation between the average question length and the accuracy of GPLLMs, both without and with CLKR (Table S4). Question lengths for MSQs and MMQs are grouped into intervals of 50 (e.g. 0–50 and 50–100), and the average question length and accuracy of seven GPLLMs are calculated for each interval (Table S4). The Pearson correlation coefficients between question length and the accuracies of GPLLMs, both with and without CLKR, are less than −0.8 (Table S4), significantly demonstrating that longer question lengths result in lower accuracy for GPLLMs. Future improvements could focus on streamlining questions to eliminate interference from irrelevant information (Saad *et al.*, 2023; Rizzo *et al.*, 2024), thereby achieving higher performance in CLQA tasks.

Besides the particular types of legal questions the LLMs struggled with, hallucination is a commonly seen reason for imperfect CLQA performance (Su *et al.*, 2024; Mansurova



(a) Power law distribution of unranked frequency of knowledge chunk-sourced documents



(b) Power law distribution of ranked frequency of knowledge chunk-sourced documents

Figure 10. The power law distribution of CL knowledge-sourced documents for CLQA. **Source(s):** Authors' own work

et al., 2024). This study employs two methods (i.e. RAG and a specifically designed prompt) to mitigate hallucinations. First, this approach provides three CL knowledge chunks to GPLLMs for CLQA; the GPLLMs refer to these CL knowledge chunks to answer questions. This method effectively reduces hallucinations by supplying relevant information from external knowledge sources (Alan *et al.*, 2024; Hou and Zhang, 2024; Mansurova *et al.*, 2024). Additionally, to minimize hallucinations during the CLQA process, a prompt is designed to inform the GPLLMs that (1) answers must be based on the provided knowledge and (2) responses should be limited to the selected option without any additional explanation. The 29,960 answers from the seven pairs of GPLLMs are presented in Table S4, where no significant issues of redundancy, irrelevance, or factual inaccuracies are observed.

5.2 Potential methods for addressing long-tail effect among CL documents

To address the long-tail effect for more efficient knowledge chunk retrieval, scholars have primarily proposed two potential methods, including (1) building cross-scale CLKRs and (2) removing redundant knowledge chunks. The documents within CLKR can be categorized into the frequently used CLKR (FU-CLKR) and rarely used CLKR (RU-CLKR) based on a certain threshold of retrieval times. During CLQA processing, the system prioritizes retrieving relevant documents from FU-CLKR. If the similarity between the question vector and the knowledge vectors in FU-CLKR falls below a certain threshold, the RAG then extends its search to knowledge chunks in the RU-CLKR (Sasazawa *et al.*, 2023). The cross-scale CLKRs

hold the potential to reduce the computational complexity of initial retrieval by focusing on a smaller, more relevant subset of CLKR (Zhu *et al.*, 2024; Sasazawa *et al.*, 2023). Secondly, although documents with similar name entities are filtered and aligned by experts during the establishment of CLKR (Figure 2b), different documents may contain overlapping knowledge, resulting in the redundancy of CLKR and complicating the retrieval process (Yu *et al.*, 2024; Gao *et al.*, 2023). The duplicated or similar contexts in different documents could be eliminated or merged (Gao *et al.*, 2023; Yu *et al.*, 2024; Zhou *et al.*, 2023), ensuring a more streamlined and efficient knowledge repository.

5.3 Practical implications of this study

The practical implications of this study include (1) offering an additional channel for CLQA query, (2) highlighting the importance of adding evolving domain-specific knowledge in GPLLM application, and (3) holding the transferability to develop region-specific CLKR beyond Mainland China.

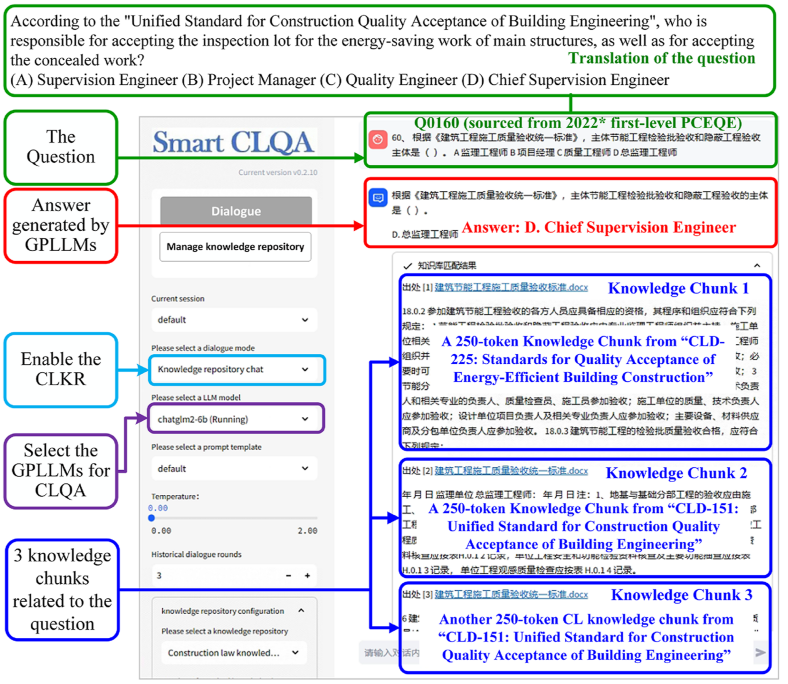
This study provides engineers with an additional channel of CLQA (Figure 11), complementing traditional ways such as consulting books (Rasool *et al.*, 2024; Choi *et al.*, 2023), conducting online searches (Oeding *et al.*, 2024), and seeking expert advice (Hou and Zhang, 2024; Alan *et al.*, 2024). Regardless of whether the questions are open-ended or closed-ended, this approach is more efficient than consulting books and online searches while being more cost-effective than seeking expert advice (Chou *et al.*, 2024; Lee *et al.*, 2023). Although current GPLLM-based CLQA systems cannot perfectly answer all questions, they serve as a meaningful supplement to traditional ways.

In practical GPLLM applications including CLQA, this study highlights the importance of incorporating domain-specific knowledge. The addition of domain-specific knowledge not only improves accuracy but also enhances the explainability of answers (Figure 11a) (Mansurova *et al.*, 2024; Su *et al.*, 2024). CLQA knowledge, in particular, is constantly evolving. New laws in the construction industry are continuously introduced, while existing laws may be modified or repealed (Li *et al.*, 2021; Tian *et al.*, 2023). The prototype developed in this study includes a feature for updating CL documents in the CLKR (Figure 11b). This facilitates the incorporation of up-to-date CL documents and the removal of outdated ones, ensuring that the knowledge base remains current and relevant.

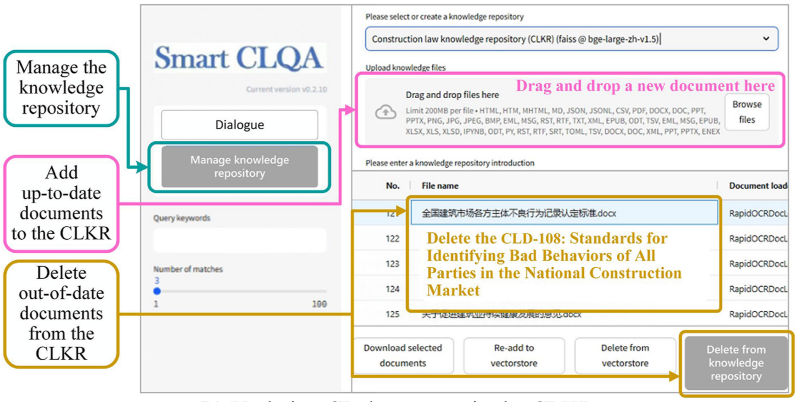
Finally, this proposed methodology can be referred to develop region-specific CLKRs beyond Mainland China. As variations in construction-related laws, regulations, and standards exist across different countries and regions (Alhyari and Ani, 2022; Hansen, 2013), scholars should collect region-specific corpora (e.g. local written judgments and textbooks) and then identify candidate legal documents automatically. While this study presents an example of a "1-8-164" three-layer knowledge hierarchy (Figure 2b), it is crucial to consult region-specific experts to construct the CL hierarchy for filtering and categorizing CL documents. By the two adjustments, the proposed data-driven and expertise-based approach for establishing external knowledge bases can be applied to regions beyond Mainland China. All related codes for collecting raw documents and building CLKR are shared in a GitHub repository.

5.4 Limitations and further endeavors

This study still shares limitations with existing literature and calls for further endeavors, including (1) the challenges of using large-scale open-ended questions in the CLQA dataset (Table 2), (2) the need to mitigate long-tail effects of CLKR, and (3) the necessity to continually update the CLKR and the selected GPLLMs. While this research includes a set of 100 open-ended questions (Table S4), performance testing on a larger-scale open-ended question set presents significant challenges, including the subjective nature of answer evaluation and the considerable time and effort required. Additionally, a long-tail effect among CL documents within CLKR is identified. Future improvements could focus on developing cross-scale CLKR (Zhu *et al.*, 2024; Sasazawa *et al.*, 2023) and removing redundant



(a) A CLQA example in the deployed prototype



(b) Updating CL documents in the CLKR

Figure 11. The CLQA prototype and the CLKR update. Note: Codes and specifications for deploying the prototype are available in supplemental materials. **Source(s):** Authors' own work

knowledge chunks (Yu *et al.*, 2024; Gao *et al.*, 2023) to streamline CLKR and maximize the contribution of individual documents. Finally, It is acknowledged that achieving perfect CLQA remains a formidable challenge for current GPLLM technology. The constant updates to the knowledge in CLKR and the continuous development of more powerful GPLLM models are crucial for enhancing the performance of CLQA in the future, which requires the collaborative efforts of scholars worldwide.

6. Conclusion

The study proposes an approach of developing CLKR to empower GPLLMs for CLQA. The development process involves (1) identifying 702 candidate documents from 374,992 pieces of written judgments (Figure 2a), (2) building a CLKR with 387 documents covering over 8 main areas and 164 subareas (Figure 2b), (3) conducting a three-step integration between CLKR and GPLLMs (Figure 3), and (4) constructing a 2,140-question validation dataset (Figure 4) to evaluate the efficacy of CLKR. The findings indicate that CLKR notably augments GPLLMs' performance in CLQA (Figure 6–9 and Tables 4–6), enhancing the accuracy by an average of 21.1%. The CLQA performance enhancements across various CL knowledge areas vary from 14.5% to 28.2% (Table 6).

Three major contributions are concluded as follows. Firstly, this study devises a data-driven and expertise-based method to construct the external knowledge base. Previous studies rely heavily on expertise to manually select documents, limiting QA to a narrow scope and rarely considering reuse and updating (Table 1). This approach not only relieves the reliance on experts during document selection but also broadens the knowledge scope of an external knowledge base aligning with the professional examinations (Figure 2). Secondly, this study empirically demonstrates the effectiveness of the CLKR in improving the performance of GPLLMs on CLQA (Figure 6–9 and Table 4–6). Contrasting with existing CLQA-related approaches (Table 1), CLKR-augmented GPLLM-based QA holds the advantages of avoiding complex training of conventional learning-based QA models and better CLQA performance. Finally, this work offers an openly available test dataset (Figure 4) as a benchmark dataset to advance the CLQA field. Previous studies have primarily relied on existing datasets and seldom provide access to the test dataset (Tables 1 and 2). This study provides a CLQA dataset comprising 2,140 questions from authoritative PCEQEs (Figure 4) to help objectively assess the effectiveness and comprehensiveness of CLKR.

References

Alan, A.Y., Karaarslan, E. and Aydin, O. (2024), "A RAG-Based question answering system proposal for understanding Islam: MufassirQAS LLM", *arXiv*, pp. 1-19, doi: 10.48550/arXiv.2401.15378.

Alhyari, O.H. and Ani, A.R.A. (2022), "Is the engineering and construction contract legally less competitive than the red book in civil law countries?", *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, Vol. 14 No. 3, 06522001, doi: 10.1061/(asce)la.1943-4170.0000543.

Antaki, F., Touma, S., Milad, D., El-Khoury, J. and Duval, R. (2023), "Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings", *Ophthalmology Science*, Vol. 3 No. 4, 100324, doi: 10.1016/j.xops.2023.100324.

Badyal, D.K., Jain, A., Lata, H. and Sharma, M. (2023), "Triple Cs of scenario-based multiple-choice question: concept, construction, and corroboration", *National Journal of Pharmacology and Therapeutics*, Vol. 1 No. 1, pp. 117-122, doi: 10.4103/njpt.njpt_47_23.

Choi, J.H., Hickman, K.E., Monahan, A.B. and Schwarcz, D.B. (2023), "ChatGPT goes to law school", *SSRN Electronic Journal*, doi: 10.2139/ssrn.4335905.

Chou, J.-S., Chong, P.-L. and Liu, C.-Y. (2024), "Deep learning-based chatbot by natural language processing for supportive risk management in river dredging projects", *Engineering Applications of Artificial Intelligence*, Vol. 131, 107744, doi: 10.1016/j.engappai.2023.107744.

deepset-ai (2024), "Haystack", available at: https://github.com/deepset-ai/haystack

DJ (2023), "A statistical analysis of lawyer and grassroots legal service work in 2022", available at: https://www.moj.gov.cn/pub/sfbgw/zwxxgk/fdzdgknr/fdzdgknrtjxx/202306/t20230614_480740.html

Domengo (2024), "PythonProject", available at: https://github.com/Domengo/pythonProject/blob/master/llm-chat/langchain_gemini_qa.py

Eleliemy, A. and Ciorba, F.M. (2021), "A distributed chunk calculation approach for self-scheduling of parallel applications on distributed-memory systems", *Journal of Computational Science*, Vol. 51, 101284, doi: 10.1016/j.jocs.2020.101284.

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P.C. and Berner, J. (2023), "Mathematical capabilities of chatgpt", *arXiv*, pp. 1-37, doi: 10.48550/arXiv.2301.13867.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. and Wang, H. (2023), "Retrieval-augmented generation for large language models: a survey", *arXiv*, pp. 1-21, doi: 10.48550/arXiv.2312.10997.

Gencer, A. and Aydin, S. (2023), "Can ChatGPT pass the thoracic surgery exam?", *The American Journal of the Medical Sciences*, Vol. 366 No. 4, pp. 291-295, doi: 10.1016/j.amjms.2023.08.001.

Ghimire, P., Kim, K. and Acharya, M. (2024), "Opportunities and challenges of generative AI in construction industry: focusing on adoption of text-based models", *Buildings*, Vol. 14 No. 1, p. 220, doi: 10.3390/buildings14010220.

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A. and Chartash, D. (2023), "How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment", *JMIR Medical Education*, Vol. 9 No. 1, e45312, doi: 10.2196/45312.

Hansen, A. (2013), "Building Codes and Regulations", In *The Canadian Encyclopedia*, available at: https://www.thecanadianencyclopedia.ca/en/article/building-codes-and-regulations (accessed 16 December 2013).

Harvel, N., Haiek, F.B., Ankolekar, A. and Brunner, D.J. (2024), "Can LLMs answer investment banking questions? Using domain-tuned functions to improve LLM performance on knowledge-intensive analytical tasks", *Proceedings of the AAAI Symposium Series*, Vol. 3 No. 1, pp. 125-133, doi: 10.1609/aaaiss.v3i1.31191.

Hou, Y. and Zhang, R. (2024), "Enhancing dietary supplement question answer via retrieval-augmented generation (RAG) with LLM", *medRxiv*, pp. 1-15, doi: 10.1101/2024.09.11.24313513.

Hwang, Y. and Mattila, A.S. (2022), "The effect of smart shopper self-perceptions on word-of-mouth behaviors in the loyalty reward program context", *Journal of Hospitality & Tourism Research*, Vol. 46 No. 2, pp. 243-266, doi: 10.1177/1096348020985212.

infiniflow (2024), "Ragflow", available at: https://github.com/infiniflow/ragflow

Khademi Adel, T., Modir, M. and Ravanshadnia, M. (2022), "An analytical review of construction law research", *Engineering Construction and Architectural Management*, Vol. 29 No. 5, pp. 1931-1945, doi: 10.1108/ecam-05-2020-0306.

Khan, I. (2024), "Microsoft's copilot embraces the power of openAI's new GPT-4o", available at: https://www.cnet.com/tech/services-and-software/microsoft-copilot-embraces-the-power-of-openais-new-gpt-4-o/

Kim, J., Chung, S. and Chi, S. (2024), "Cross-lingual information retrieval from multilingual construction documents using pretrained language models", *Journal of Construction Engineering and Management*, Vol. 150 No. 6, 04024041, doi: 10.1061/jcemd4.coeng-14273.

Kortemeyer, G. (2023), "Could an artificial-intelligence agent pass an introductory physics course?", *Physical Review Physics Education Research*, Vol. 19 No. 1, 010132, doi: 10.1103/physrevphyseducres.19.010132.

Langchain-ai (2024), "Langchain", available at: https://github.com/langchain-ai/langchain

Lee, S.-H., Choi, S.-W. and Lee, E.-B. (2023), "A question-answering model based on knowledge graphs for the general provisions of equipment purchase orders for steel plants maintenance", *Electronics*, Vol. 12 No. 11, p. 2504, doi: 10.3390/electronics12112504.

Li, H., He, L. and Zeng, H. (2021), *Construction Engineering Regulations*, Nanjing University Press, Nanjing.

Liu, J.Y. and Low, S.P. (2011), "Work–family conflicts experienced by project managers in the Chinese construction industry", *International Journal of Project Management*, Vol. 29 No. 2, pp. 117-128, doi: 10.1016/j.ijproman.2010.01.012.

Liu, H., Li, J., Li, H., Li, H., Mao, P. and Yuan, J. (2021), "Risk perception and coping behavior of construction workers on occupational health risks—A case study of Nanjing, China", *International Journal of Environmental Research and Public Health*, Vol. 18 No. 13, p. 7040, doi: 10.3390/ijerph18137040.

Liu, L., Meng, J. and Yang, Y. (2024), "LLM technologies and information search", *Journal of Economy and Technology*, Vol. 2, pp. 269-277, doi: 10.1016/j.ject.2024.08.007.

Lu, J., Tian, X., Zhang, C., Zhao, Y., Zhang, J., Zhang, W., Feng, C., He, J., Wang, J. and He, F. (2024 In press), "Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry", *Energy and Built Environment*, doi: 10.1016/j.enbenv.2024.03.010.

Mansurova, A., Mansurova, A. and Nugumanova, A. (2024), "QA-RAG: exploring LLM reliance on external knowledge", *Big Data and Cognitive Computing*, Vol. 8 No. 9, p. 115, doi: 10.3390/bdcc8090115.

Martinez-Gil, J. (2023), "A survey on legal question–answering systems", *Computer Science Review*, Vol. 48, 100552, doi: 10.1016/j.cosrev.2023.100552.

meta-llama (2024), "Llama2", available at: https://huggingface.co/docs/transformers/main/model_doc/llama2

neuml (2024), "Txtai", available at: https://github.com/neuml/txtai

NPC (2024), "National legal database", available at: https://flk.npc.gov.cn/index.html

Oeding, J.F., Lu, A.Z., Mazzucco, M., Fu, M.C., Taylor, S.A., Dines, D.M., Warren, R.F., Gulotta, L.V., Dines, J.S. and Kunze, K.N. (2024), "ChatGPT-4 performs clinical information retrieval tasks using consistently more trustworthy resources than does Google search for queries concerning the latarjet procedure", *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, Vol. 41 No. 3, pp. 588-597, doi: 10.1016/j.arthro.2024.05.025.

Oh, N., Choi, G.-S. and Lee, W.Y. (2023), "ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models", *Annals of Surgical Treatment and Research*, Vol. 104 No. 5, p. 269, doi: 10.4174/astr.2023.104.5.269.

pathwaycom (2024), "llm-app", available at: https://github.com/pathwaycom/llm-app

Pei, D., He, J., Liu, K., Chen, M. and Zhang, S. (2024), "Application of large language models and assessment of their ship-handling theory knowledge and skills for connected maritime autonomous surface ships", *Mathematics*, Vol. 12 No. 15, p. 2381, doi: 10.3390/math12152381.

Petrus, S. (2024), "Top 10 RAG frameworks Github Repos 2024", available at: https://sebastian-petrus.medium.com/top-10-rag-frameworks-github-repos-2024-12b2a81f4a49

Pursnani, V., Sermet, Y., Kurt, M. and Demir, I. (2023), "Performance of ChatGPT on the US fundamentals of engineering exam: comprehensive assessment of proficiency and potential implications for professional environmental engineering practice", *Computers and Education: Artificial Intelligence*, Vol. 5, 100183, doi: 10.1016/j.caeai.2023.100183.

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016), "Squad: 100,000+ questions for machine comprehension of text", *arXiv*, pp. 1-10, doi: 10.48550/arXiv.1606.05250.

Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., Hampstead, B.M., Belleville, S., Mouzakis, K. and Bahar-Fuchs, A. (2024), "Evaluating LLMs on document-based QA: exact answer selection and numerical extraction using CogTale dataset", *Natural Language Processing Journal*, Vol. 8, 100083, doi: 10.1016/j.nlp.2024.100083.

Rizzo, M.G., Cai, N. and Constantinescu, D. (2024), "The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education", *Journal of Orthopaedics*, Vol. 50, pp. 70-75, doi: 10.1016/j.jor.2023.11.056.

Rosól, M., Gąsior, J.S., Łaba, J., Korzeniewski, K. and Młyńczak, M. (2023), "Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination", *Scientific Reports*, Vol. 13 No. 1, 20512, doi: 10.1038/s41598-023-46995-z.

RUC-NLPIR (2024), "FlashRAG", available at: https://github.com/RUC-NLPIR/FlashRAG

Saad, A., Iyengar, K.P., Kurisunkal, V. and Botchu, R. (2023), "Assessing chatgpt's ability to pass the FRCS orthopaedic part A exam: a critical analysis", *The Surgeon*, Vol. 21 No. 5, pp. 263-266, doi: 10.1016/j.surge.2023.07.001.

Sahin, M.C., Sozer, A., Kuzucu, P., Turkmen, T., Sahin, M.B., Sozer, E., Tufek, O.Y., Nernekli, K., Emmez, H. and Celtikci, E. (2024), "Beyond human in neurosurgical exams: chatgpt's success in the Turkish neurosurgical society proficiency board exams", *Computers in Biology and Medicine*, Vol. 169, 107807, doi: 10.1016/j.compbiomed.2023.107807.

Saka, A., Taiwo, R., Saka, N., Salami, B.A., Ajayi, S., Akande, K. and Kazemi, H. (2024), "GPT models in construction industry: opportunities, limitations, and a use case validation", *Developments in the Built Environment*, Vol. 17, 100300, doi: 10.1016/j.dibe.2023.100300.

Sasazawa, Y., Yokote, K., Imaichi, O. and Sogawa, Y. (2023), "Text retrieval with multi-stage Re-Ranking models", *arXiv*, pp. 1-7, doi: 10.48550/arXiv.2311.07994.

Schoch, J., Schmelz, H.U., Borgmann, H. and Nestler, T. (2024), "A0179 - performance of ChatGPT on the fellow of the European Board of Urology (FEBU) exams: a comparative analysis", *European Urology*, Vol. 85, pp. S923-S924, doi: 10.1016/s0302-2838(24)00759-0.

Shi, Y., Wang, J., Ren, P., ValizadehAslani, T., Zhang, Y., Hu, M. and Liang, H. (2023), "Fine-tuning BERT for automatic ADME semantic labeling in FDA drug labeling to enhance product-specific guidance assessment", *Journal of Biomedical Informatics*, Vol. 138, 104285, doi: 10.1016/j.jbi.2023.104285.

SPC (2023), "The total number of cases accepted by courts nationwide exceeded 33 million in 2022", available at: https://www.chinacourt.org/article/detail/2023/03/id/7178559.shtml

Su, M.-C., Lin, L.-E., Lin, L.-H. and Chen, Y.-C. (2024), "Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: insights from Taiwan's nursing licensing exam", *International Journal of Nursing Studies*, Vol. 153, 104717, doi: 10.1016/j.ijnurstu.2024.104717.

Sun, J., Lei, K., Cao, L., Zhong, B., Wei, Y., Li, J. and Yang, Z. (2020), "Text visualization for construction document information management", *Automation in Construction*, Vol. 111, 103048, doi: 10.1016/j.autcon.2019.103048.

Tao, G., Feng, J., Feng, H., Feng, H. and Zhang, K. (2022), "Reducing construction dust pollution by planning construction site layout", *Buildings*, Vol. 12 No. 5, p. 531, doi: 10.3390/buildings12050531.

Tian, D., Li, M., Ren, Q., Zhang, X., Han, S. and Shen, Y. (2023), "Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining", *Automation in Construction*, Vol. 145, 104670, doi: 10.1016/j.autcon.2022.104670.

truefoundry (2024), "Cognita", available at: https://github.com/truefoundry/cognita

Tsoutsanis, P. and Tsoutsanis, A. (2024), "Evaluation of large language model performance on the multi-specialty recruitment assessment (MSRA) exam", *Computers in Biology and Medicine*, Vol. 168, 107794, doi: 10.1016/j.compbiomed.2023.107794.

Wang, X. and El-Gohary, N. (2023), "Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements", *Automation in Construction*, Vol. 147, 104696, doi: 10.1016/j.autcon.2022.104696.

Wang, K., Xiong, L., Liu, A., Zhang, G. and Lu, J. (2024), "A self-adaptive ensemble for user interest drift learning", *Neurocomputing*, Vol. 577, 127308, doi: 10.1016/j.neucom.2024.127308.

Xu, N., Liang, Y., Guo, C., Meng, B., Zhou, X., Hu, Y. and Zhang, B. (2023), "Entity recognition in the field of coal mine construction safety based on a pre-training language model", *Engineering Construction and Architectural Management*, Vol. 32 No. 4, pp. 2590-2613, doi: 10.1108/ecam-05-2023-0512.

Xue, X., Zhang, J. and Chen, Y. (2024), "Question-answering framework for building codes using fine-tuned and distilled pre-trained transformer models", *Automation in Construction*, Vol. 168, 105730, doi: 10.1016/j.autcon.2024.105730.

Yu, L., Liu, B., Lin, Q., Zhao, X. and Che, C. (2024), "Semantic similarity matching for patent documents using ensemble BERT-Related model and novel text processing method", *arXiv*, pp. 1-5, doi: 10.48550/arXiv.2401.06782.

Zailani, B., Moda, H., Ibrahim, Y. and Abubakar, M. (2023), "Improving the antecedents of non-compliance to safety regulations toward an optimized self-regulated construction environment in Nigeria", *International Journal of Occupational Safety and Ergonomics*, Vol. 29 No. 3, pp. 1212-1219, doi: 10.1080/10803548.2022.2115657.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D. and Xing, E. (2023), "Judging llm-as-a-judge with mt-bench and chatbot arena", *Advances in Neural Information Processing Systems*, Vol. 36, pp. 46595-46623.

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T. and Fang, W. (2020), "Deep learning-based extraction of construction procedural constraints from construction regulations", *Advanced Engineering Informatics*, Vol. 43, 101003, doi: 10.1016/j.aei.2019.101003.

Zhou, H., Gao, B., Tang, S., Li, B. and Wang, S. (2023), "Intelligent detection on construction project contract missing clauses based on deep learning and NLP", *Engineering Construction and Architectural Management*, Vol. 32 No. 3, pp. 1546-1580, doi: 10.1108/ecam-02-2023-0172.

Zhu, P., Wang, Z., Okumura, M. and Yang, J. (2024), "MRHF: multi-stage retrieval and hierarchical fusion for textbook question answering", *International Conference on Multimedia Modeling*, pp. 98-111, doi: 10.1007/978-3-031-53308-2_8.

Supplementary material
The supplementary material for this article can be found online.

Corresponding author
Dezhi Li can be contacted at: njldz@seu.edu.cn