

Full length article

Augmenting general-purpose large-language models with domain-specific multimodal knowledge graph for question-answering in construction project management



Shenghua Zhou^{a,b}, Keyan Liu^a, Dezhi Li^{a,*}, Chun Fu^c, Yan Ning^d, Wenying Ji^e, Xuefan Liu^a, Bo Xiao^f, Ran Wei^b

^a China-Pakistan Belt and Road Joint Laboratory on Smart Disaster Prevention of Major Infrastructures, Southeast University, Nanjing 211189, China

^b Department of Engineering, University of Cambridge, Cambridge CB3 0FA, UK

^c Department of Engineering Management, Central South University, Changsha 410083, China

^d School of Management & Engineering, Nanjing University, Nanjing 210008, China

^e Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, VA, USA

^f Department of Civil, Environmental, and Geospatial Engineering, Michigan Technological University, Houghton, MI, USA

ARTICLE INFO

Keywords:

GLM
Construction Project Management
QA
Multimodal Knowledge Graph

ABSTRACT

Current studies on Question-Answering of Construction Project Management (CPM-QA) face challenges, including the small-scale CPM-related knowledge repositories, the limited effectiveness of QA methods using grammar rules or tiny machine-learning models, and the shortage of testing sets for comparing QA performance. Hence, this research augments general-purpose large-language models (GLMs) with the multimodal CPM knowledge graph (CPM-KG) for CPM-QA. It encompasses (i) building the multimodal CPM-KG covering 36 CPM subfields, (ii) combining CPM-KG and GLMs through three stages, (iii) developing a 2435-question CPM-QA testing set, and (iv) assessing and comparing CPM-QA accuracies for eight pairs of original and CPM-KG-augmented GLMs. The results demonstrate that CPM-KG-augmented GLMs' CPM-QA accuracy rate is 30.0 % superior to original GLMs on average, and top-performing CPM-KG-augmented GLMs (e.g., ERNIE-Bot 4.0) pass CRCEEs. Within 36 CPM subfields, CPM-QA accuracy enhancements resulting from CPM-KG are between 12.2 % and 57.8 %. Furthermore, CPM-KG leads to CPM-QA accuracy enhancements of 19.6 % for single-answer, 48.0 % for multiple-answer, 30.6 % for text-only, and 20.4 % for image-embedded questions. The multimodal CPM-KG also outperforms the text-only single-modal CPM-KG in enhancing CPM-QA performance. This work contributes to unveiling the significance of CPM-specific knowledge in augmenting GLMs, sharing a reusable multimodal CPM-KG-formatted knowledge repository, and delivering a testing set of CPM-QA.

1. Introduction

Construction Project Management (CPM) covers various subfields such as organization, schedule, cost, environment, and information management [1,2], driving the worldwide construction market worth USD 13.57 trillion by 2023 [3]. When meeting CPM-related questions, constructors and engineers typically rely on consulting literature (e.g., books, articles, and reports), using internet information, or seeking advice from CPM experts. However, reviewing literature is time-consuming, as it requires sifting through vast amounts of text to find

specific answers [4]. The reliance on internet information often results in inaccuracies and inefficiencies, as the internet search tools (e.g., Bing) only provide multiple pieces of question-related information rather than giving direct answers to specific CPM questions [5]. While consulting domain experts is effective, it is impeded by considerable consultation costs and the limited availability of CPM authoritative experts [6]. Hence, smart CPM question-answering (CPM-QA) emerges as a promising supplement, offering a highly-efficient, cost-effective, and easily-accessible way of providing referable answers to CPM questions [7,8].

Most current research on intelligent CPM-QA relies on grammar rule-

* Corresponding author at: Southeast University, Room 1218, Civil Engineering Building, School of Civil Engineering, Southeast University, Jiangning District, Nanjing 211189, China.

E-mail addresses: shenghua@connect.hku.hk (S. Zhou), lky9858@163.com (K. Liu), njldz@seu.edu.cn (D. Li), fuchun@mail.csu.edu.cn (C. Fu), ny@nju.edu.cn (Y. Ning), wji2@gmu.edu (W. Ji), 03203507@cumt.edu.cn (X. Liu), boxiao@mtu.edu (B. Xiao), rw741@cam.ac.uk (R. Wei).

based retrieval methods [9,10] or tiny machine-learning models [11,12]. In the grammar rule-based CPM-QA studies, a wide spectrum of rules concerning indicative terms, lemmas, and dependency trees is devised to search for answers in predefined CPM-related knowledge repositories (e.g., text files, ontologies, and SQL databases) [13,14]. The rule-based CPM-QA usually underperforms because of limitations regarding the knowledge repository and the rule design [15,16]. Specifically, the predefined knowledge repositories contain limited CPM knowledge and concentrate only on text modality [10], and the complexity of grammatical rules also makes it challenging to develop comprehensive rules that fully cover the wide scope of CPM-QA [17]. Furthermore, current studies employ small-scale machine-learning models for CPM-QA, such as Conditional Random Field model [18], BiLSTM [19], and BERT [11,20]. Traditional machine-learning models, with their smaller parameter sizes, often struggle to comprehend context and have limited language processing ability [21,22]. No matter the CPM-QA based on grammar rules or machine-learning models, cutting-edge studies are only capable of achieving CPM-QA in a limited scope (e.g., QA within one CPM-related document), which is far from the intelligent QA across dozens of CPM subfields (e.g., quality, schedule, cost, and contract).

The state-of-the-art developments in natural language processing (NLP) have been characterized by general-purpose large-language models (GLMs) such as Llama, GPT, and ChatGLM [23]. Such GLMs have significantly surpassed the rule-based and machine-learning models in handling a wide spectrum of NLP tasks, benefiting from their sophisticated contextual understanding capabilities [24,25]. Nevertheless, GLMs lack CPM-specific knowledge in multiple modalities and are unable to update CPM knowledge after development [26]. Consequently, the GLMs may underperform in CPM-QA [27], suggesting the infusion of CPM-specific knowledge to augment GLMs' CPM-QA capabilities [28].

This study advances CPM-QA by incorporating the CPM knowledge graph (CPM-KG) into GLMs. The process unfolds in four phases: (i) building a multimodal CPM-KG covering 36 CPM subfields; (ii) formulating a three-stage CPM-KG and GLM integration process via Retrieval-Augmented Generation (RAG); (iii) devising a 2,435-question CPM-QA testing set by collecting all questions in Chinese Registered Construction Engineer Examinations (CRCEEs) during 2013–2023; and (iv) assessing and comparing CPM-QA accuracies between original GLMs and CPM-KG-augmented GLMs. With the CPM-QA of Mainland China as a demonstration, the research results show significant CPM-QA performance enhancements and underscore the significance of CPM-specific knowledge for CPM-QA of GLMs. Additionally, this study shares a developed CPM-KG and a CPM-QA testing set, both of which are cornerstones in the CPM-QA research area to push the development of smart CPM-QA.

2. Literature review of related works

2.1. CPM-related QA studies

Many scholars have conducted CPM-related QA studies, as illustrated in Table 1. These studies involve incident query [11], compliance checking [9,12], construction information retrieval [10,19], construction behaviour [14], quality management [13], and contract management [18]. While current studies focus on diverse CPM-related topics, their critical elements (i.e., QA-retrieved knowledge repositories, QA-utilized NLP approaches, and assessments of QA performance) are analyzed and compared to reveal existing gaps in CPM-QA research (Table 1).

Regarding QA-retrieved knowledge repositories (Table 1), current works primarily retrieve information from various forms of knowledge repositories, such as task-specific documents (e.g., IFC files) [29] and ontologies [9,10]. However, these QA-retrieved knowledge repositories have two drawbacks: (i) the limitation of a single textual modality and (ii) the limited scale of CPM-related knowledge. Most existing

knowledge repositories contain textual information only, excluding image-formatted data related to CPM (e.g., activity on node diagrams) [14,20]. Furthermore, these knowledge repositories (Table 1) are typically constrained in scale, ranging from one to several documents [13,11]. In general, current knowledge repositories are inadequate for supporting CPM-QA, as they fail to cover multiple CPM subfields and the extensive range of CPM-related documents required.

Existing CPM-related QA studies predominantly use grammar rule-based approaches and small-scale machine-learning models (Table 1). Grammar rule-based approaches employ manually crafted rules concerning keywords, indicator terms, parts of speech, lemmas, dependency trees, and syntactic structures to derive answers from knowledge repositories [30,31]. Although they require no training, these methods are often critiqued for their restricted generative capacity [32,33]. When dealing with complex QA tasks, rule-based methods may encounter a phenomenon known as “rule explosion” [34]. Conversely, small-scale machine-learning models (e.g., BiLSTM and BERT) outperform grammar rule-based approaches [35]. Nevertheless, because of restrictions on their parameter sizes [22], machine-learning models still have limited language processing abilities [21], which poses challenges in addressing CPM-QA.

In the stage of retrieving question-related knowledge, current CPM-related QA studies employ various methods (Table 2), such as embedding-based retrieval, keyword-based retrieval, and others. Each method has its specific advantages and disadvantages, along with appropriate scenarios (Table 2). When selecting a method, primary considerations include preparation requirements, retrieval speed, and retrieval cost [36]. Methods that require extensive foundational work are less favorable. For example, keyword-based retrieval requires an appropriate keyword listWang and Lo [37], and rule-based methods necessitate the design of a set of rules [38,39]. Conversely, embedding-based methods and GLM-based methods generally do not require extensive pre-retrieval preparations [40,41]. Regarding retrieval speed, most methods operate quickly; however, GLM-based retrieval methods are considerably slower because the GLM must compare the input question with a large number (possibly thousands) of knowledge pieces in the knowledge base one by one. In other words, retrieving knowledge for a single question may require running the GLM thousands of times, which amplifies the inherent slowness of GLMs [42]. From a cost perspective, most methods are free of financial charges, whereas the costs of GLM-based retrieval can be substantial. Embedding-based retrieval supports linguistic diversity, has fast processing speeds, and operates at a low cost [43,44]. Consequently, embedding-based retrieval is well-suited for extracting knowledge from large-scale knowledge bases and is appropriate for this research.

As for QA performance assessment, a persistent challenge is the shortage of publicly accessible CPM-QA testing sets [45]. Researchers have developed a variety of narrowly focused QA testing sets aimed at validating the efficacy of their smart QA works (Table 1) [46]. However, most of the QA testing sets are not readily available, making current CPM-related QA studies isolated from each other. This further prevents subsequent scholars from conducting in-depth investigations based on prior work and comparing performances using the same QA testing sets [47]. The progression of sophisticated CPM-QA necessitates openly accessible QA testing sets as benchmarks, which are essential for performance assessment and comparison.

2.2. GLM-based applications

The advent of GLMs has significantly transformed NLP tasks [48]. This transformation is attributable to the extensive parameterization of GLMs, enabling superior language understanding and processing compared to earlier, smaller-scale machine-learning models [26]. As presented in Table 3, there are a variety of GLM applications across fields such as writing [49], biology [50], education [51], medicine [52], and language [53]. Notable examples include ProGen, BloombergGPT,

Table 1
CPM-QA related studies.

No.	Publications	Research field	QA-retrieved knowledge repositories				QA-used NLP approaches			Assessments of QA performance	
			Data modality	Contents	Storage forms	Openly available	Types	Retrieval methods	Implementation	Test set	Easily available
3	[9]	Compliance checking for fire safety	Only text	Fire safety-related requirements (e.g., General Building Regulation)	Ontology	×	Rule-based method	Hierarchical or graph relationship-based knowledge retrieval	SPARQL-based rules	4 explicit implementation instances	×
	[13]	Quality compliance checking	Only text	2 specifications (SL 677–2014 & SL 632–2012) & dozens of quality record documents	Text files	×	Rule-based method & Conventional machine-learning model	Keyword-based retrieval	Syntax-related rules, BERT model, BiLSTM-CRF model	36 quality records from the Xiangjiaba Dam project	×
	[10]	BIM query	Only text	121 classes, 58 object properties, and 446 individuals	Ontology & IFC file	×	Rule-based method	Rule-based matching	MOP-SP	225 natural language queries	×
	[14]	Information retrieval of transportation construction	Only text	CLEAR entries, NCDOT construction specifications and manuals, contract documents, and construction textbooks	CLEAR database	×	Rule-based method & Conventional machine-learning model	Embedding-based retrieval	CD-SAILmodel	A subset of 5 search inputs	×
	[11]	Classification-based QA of construction incident reports	Only text	Injury cases from the OSHA dataset for model training	Structured dataset	×	Conventional machine-learning model	Embedding-based retrieval	Fine-tuned BERT model	1,169 injury cases from the OSHA dataset	×
	[75]	Construction quality compliance checking	Only text	3 common types of construction quality constraints	Ontology	×	Rule-based method	Rule-based matching	Ontology-based information extraction method	5 mandatory construction regulations	×
	[12]	Intelligent building codes	Only text	Sentences from the International Building Code and its amendments	Text files	×	Rule-based method & Conventional machine-learning model	Rule-based matching	Multi-form semantic (MFS) requirement hierarchy, RNN-based model	3 types of regulatory documents	×
	[29]	BIM query answering	Only text	Natural language paragraphs generated from BIMs for model training	IFC files	×	Conventional machine-learning model	Embedding-based retrieval	BERT model	100 pieces of test data (i.e., building attribute information-related queries)	×
	[18]	Query of change orders	Only text	1,320 change orders issued by a state roadway project	Text files	×	Rule-based method & Conventional machine-learning model	Keyword-based retrieval	Lemma, Part-of-Speech, dependency tree, and CRF	120 change orders	×
	[19]	Construction accident information extraction	Only text	Accident news reports	Text files	×	Conventional machine-learning model	Embedding-based retrieval	BiLSTM-CRF model	145 construction accident news reports	×
	[20]	Bridge detection information extraction	Only text	Bridge inspection reports in different regions of China	Text files	×	Conventional deep learning model	Embedding-based retrieval	BERT-MRC model	1,311 sentences from bridge inspection reports	×
	This work	CPM-QA	Text and image data	195 CPM-related documents with around 2.6 million words covering 36 CPM subfields	A multimodal knowledge graph named CPM-KG	√	Integration between GLMs and CPM-KG	Embedding-based retrieval	8 cutting-edge GLMs (i.e., GPT-4-turbo, ERNIE, etc.)	2,435 multiple-choice questions involving 36 CPM subfields	√

Note: SPARQL: SPARQL protocol and RDF query language; SL 677–2014: specifications for hydraulic concrete construction; SL 632–2012: inspection and assessment standard for separated item project construction quality of water conservancy and hydroelectric engineering – concrete works; Bi-LSTM: bi-directional LSTM; IFC: industry foundation classes; MOP-SP: model-based ontology population and semantic parsing; CLEAR: communicate lessons, exchange advice, record; NCDOT: north carolina department of transportation; CD-SAIL: construction domain-specific artificial intelligence language; OSHA: occupational safety and health administration; RNN: recurrent neural network; MRC: machine reading comprehension.

Table 2

Comparison of question-related knowledge retrieval methods.

Method	The procedure of the method	Benefits	Shortcomings	Applicable examples
Embedding-based retrieval	<ul style="list-style-type: none"> Generating embedding vectors for the text Computing similarity between embedding vectors 	<ul style="list-style-type: none"> Supporting linguistic diversity Processing quickly and at a low cost 	<ul style="list-style-type: none"> Requiring high-quality embeddings Poor interpretability 	<ul style="list-style-type: none"> Retrieving legal documents to find relevant legal explanations for specific cases (Su et al. 2023). Retrieving valid information from large-scale agricultural documents Peng and Yang [76]
Keyword-based retrieval	<ul style="list-style-type: none"> Determining the keywords Extracting relevant documents by using keywords 	<ul style="list-style-type: none"> Being easy to use Rapid processing and low cost 	<ul style="list-style-type: none"> High dependency on the keyword list Missing relevant information potentially 	<ul style="list-style-type: none"> Searching and analysing COVID-19 related symptoms mentioned in user self-reports using keywords Wang and Lo [37]. Retrieving electronic medical records based on fuzzy keywords (e.g., cough, insomnia) [77]. Extracting additional descriptions of spatial entities in the text regulations on underground utility infrastructure by matching adjectives, adverbs, and other modifiers Xu and Cai [38].
Rule-based matching	<ul style="list-style-type: none"> Defining matching rules Applying rules for data filtering 	<ul style="list-style-type: none"> Being fast and highly customizable No need for large amounts of training data 	<ul style="list-style-type: none"> Being rigid Lacking flexibility in handling semantic variations 	<ul style="list-style-type: none"> Building suffix matching (e.g., “.tor”, “.room”) rules to identify entities in mechanical, electrical, and plumbing fields Wu et al. [39]. Building a graph containing target nodes, source nodes, bridge nodes, attribute nodes, and relationships for financial market prediction [78]. Predicting potential interactions or adverse reactions between drugs based on the triplet relationships in the single-modal graph [79].
Hierarchical or graph relationship-based knowledge retrieval	<ul style="list-style-type: none"> Building relationships of knowledge Querying knowledge with relationship 	<ul style="list-style-type: none"> Allowing multi-level searching Optimizing knowledge storage 	<ul style="list-style-type: none"> Requiring substantial initial work of building relationships of knowledge Difficulty with high-dimensional storage 	<ul style="list-style-type: none"> Using the smaller open-source Med42-70B model to retrieve and answer clinical questions [41]. Using the Program of Thought Prompting method to generate search code for guiding the GLMs to retrieve information from the knowledge base [42].
GLM-based retrieval	<ul style="list-style-type: none"> Inputting query to the GLMs Comparing the query with each piece of knowledge in the knowledge base one by one 	<ul style="list-style-type: none"> Providing comprehensive context understanding 	<ul style="list-style-type: none"> May not act as expected Slow speed Requiring high computing resource 	

Table 3

Examples of GLM-based research and applications.

No.	Source	Research field	Research purpose	Involved GLMs	GLM performance assessment
1	[51]	Mathematics	Test the mathematical ability of GLMs	ChatGPT & GPT-4	GHOSTS dataset
2	[80]	Finance	Have a GLM focused on the financial field	BloombergGPT	2 types of tasks (finance-specific & general-purpose tasks)
3	[81]	Clinical medicine	Improve clinical knowledge assessment for GLM	PaLM & Flan-PaLM	MultiMedQA combining 6 existing medical QA datasets
4	[52]	Medical QA	Enhance the ability to retrieve medical knowledge, reason about it, and answer medical questions	Med-PaLM 2	1066 consumer medical questions and 240 long-form “adversarial” questions
5	[82]	Chemical physics	Present a GLM-based Intelligent Agent (IA) system to achieve autonomous scientific research	GPT-3.5 & GPT-4	3 distinct examples
6	[83]	Chemistry	Investigate ChatGPT’s function in answering chemical assessment questions	ChatGPT	2 chemistry-focused modules
7	[50]	Biotechnology	Develop GLM to generate artificial proteins across multiple families and functions	ProGen	100 sequences
8	[49]	Story writing	Develop Wordcraft (a web application) to allow people to collaborate with GLM on stories	LaMDA	25 hobbyist writers to write short stories using Wordcraft
9	[53]	Language	Provide models, massive high-quality corpora, and evaluation sets for the Spanish language	MarIA (a family of Spanish language models)	9 existing evaluation datasets and 1 novel extractive dataset
10	This study	Construction project management	Augment GLMs on CPM-QA	8 GLMs (i.e., GPT-4-turbo, ERNIE etc.)	2,435 CPM-related multiple-choice questions

and PaLM (Table 3). Although the works have distinct names (Table 3), they are developed on GLMs, such as Llama2 and ChatGPT. The main limitation of the GLMs is their deficiency in specific domain knowledge [24], which compromises their effectiveness in CPM-QA.

2.3. Gaps in current works

Previous works have significantly contributed to intelligent CPM-QA, laying a solid groundwork for this study. However, most current CPM-related QA investigations (Table 1) encounter three major limitations: (i) limited-scale text-only CPM-related knowledge repositories for CPM-QA retrieval, (ii) poor performance of CPM-QA methods that use grammar rules or machine-learning models, and (iii) the shortage of comparable CPM-QA performance testing sets. To bridge the three gaps, the CPM-QA requires a multimodal CPM knowledge repository covering dozens of CPM subfields, the substitution of traditional NLP methods with GLMs, and the freely accessible large-scale CPM-QA testing set.

Particularly, the CPM-specific knowledge repository would also address the deficiency of CPM knowledge in GLMs.

3. Methodology

The methodology of leveraging the multimodal CPM-KG to augment CPM-QA of GLMs is devised as displayed in Fig. 1. It comprises four primary phases: (i) the multimodal CPM-KG development regarding 36 CPM subfields; (ii) the three-stage RAG-based integration between GLMs and CPM-KG; (iii) the creation of a 2,435-question CPM-QA testing set; and (iv) the assessments and comparisons of CPM-QA performance of eight pairs of GLMs adopting and omitting CPM-KG. Given the pronounced regional differences in CPM, developing a globally applicable CPM-KG for CPM-QA is challenging, if not infeasible. Therefore, the methodology is exemplified through CPM-QA within the context of China.

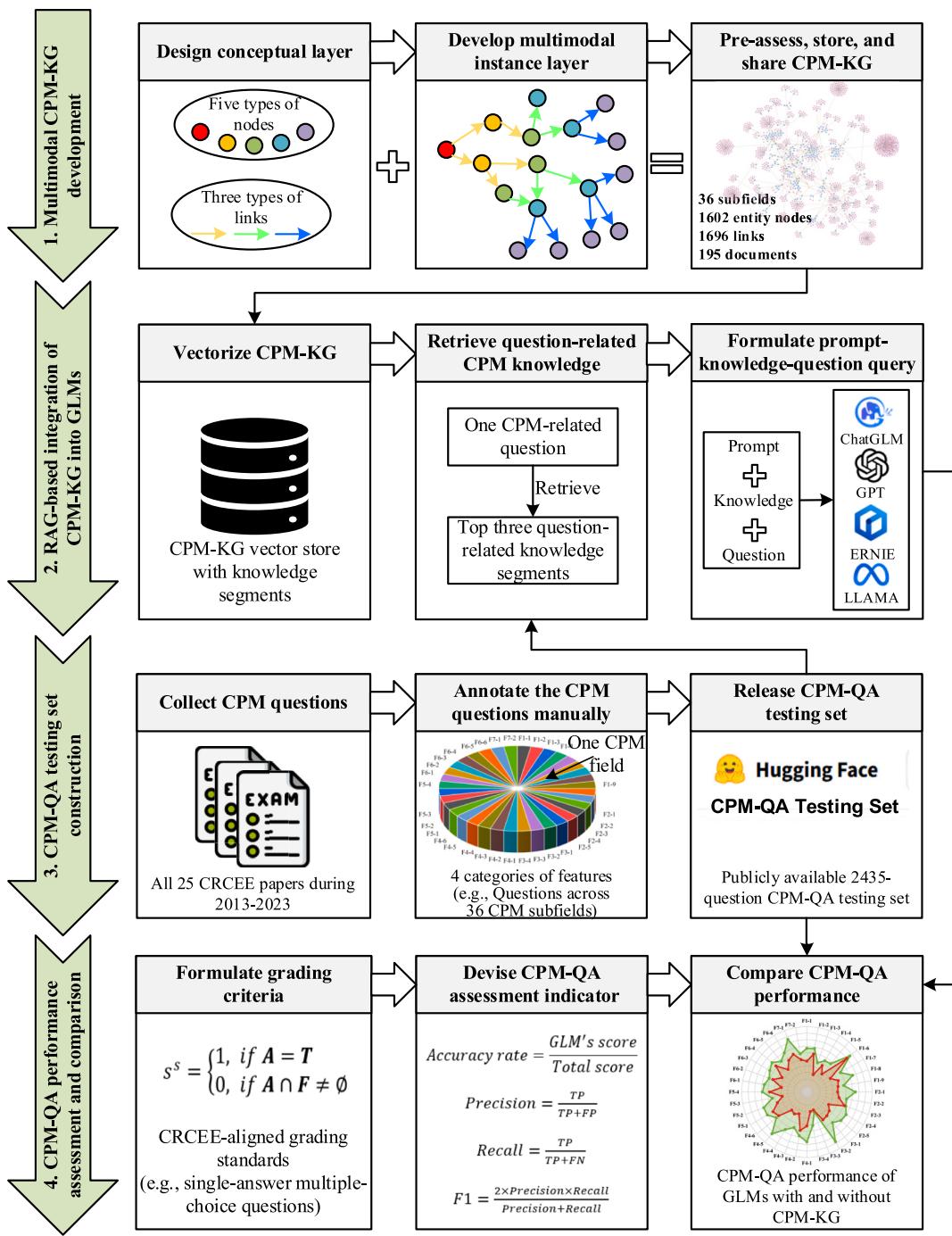


Fig. 1. Phases of leveraging the multimodal CPM-KG to augment CPM-QA of GLMs.

3.1. Multimodal CPM-KG development

As displayed in Fig. 1, developing CPM-KG consists of (i) the design of the conceptual layer, (ii) the establishment of an instance layer with multimodal data, and (iii) the pre-assessment, storage, and sharing of the CPM-KG. The details of the three steps are depicted in Fig. 2.

3.1.1. Conceptual layer development of the CPM-KG

Designing the conceptual layer of CPM-KG (Fig. 2a) involves defining the types of critical concepts (i.e., nodes) and clarifying the categories of relationships (i.e., links). The conceptual layer defines five types of nodes: primary knowledge field, secondary knowledge field, tertiary knowledge field, document, and document content (Fig. 2a) [54,55,56].

The primary, secondary, and tertiary knowledge fields refer to different tiers of CPM knowledge subfields. “Document” pertains to the materials encapsulating CPM knowledge. “Document content” means specific chapters and sections in these documents. Three types of relationships are delineated: “has subfield of”, “involves”, and “contains” (Fig. 2a). These concepts and relationships constitute four types of triplets: [primary knowledge field, has subfield of, secondary knowledge field], [secondary knowledge field, has subfield of, tertiary knowledge field], [tertiary knowledge field, involves, document], and [document, contains, document content] (Fig. 2a).

3.1.2. Establishment of CPM-KG instance layer with multimodal data

Developing the CPM-KG instance layer (Fig. 2b) involves building

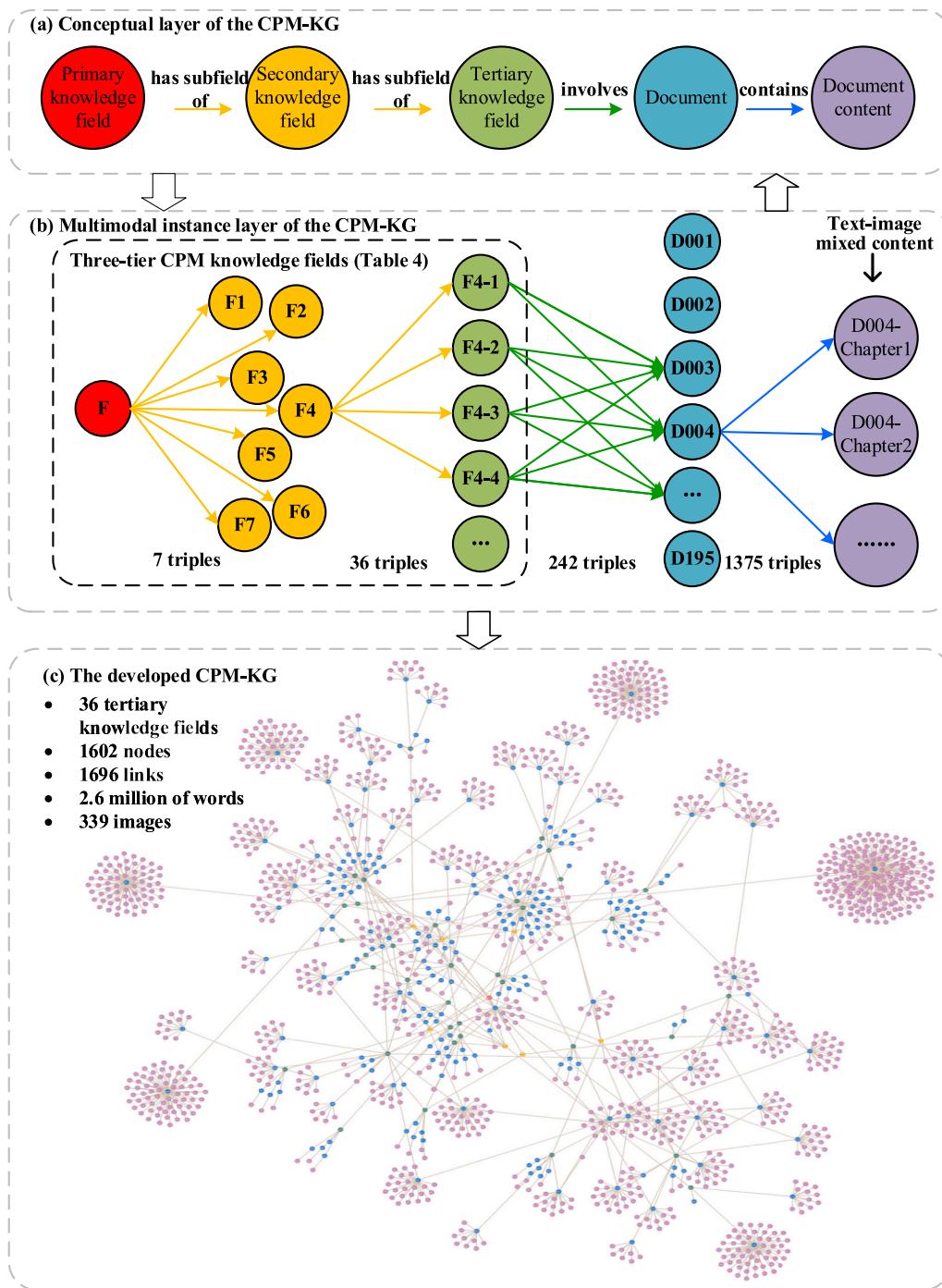


Fig. 2. CPM-KG development.

the three-level architecture of CPM knowledge fields, collecting CPM-related documents, and processing multimodal document content including text and image data.

The first step involves developing a three-level architecture for the CPM knowledge fields (Fig. 3), referencing three widely used CPM textbooks: Textbook I from China Architecture Industry Press, Textbook II from Southeast University Press, and Textbook III from Wuhan University Press [54,55,56]. Given the slight differences in the knowledge architectures of these textbooks, five experts in the CPM field (Table S1) are consulted to derive the CPM knowledge architecture. Initially, Textbooks I, II, and III each comprise 7, 8, and 7 chapters, respectively. The first chapter of Textbook II, titled “B1 Introduction to Construction Project Management”, is not treated as a standalone chapter in

Textbooks I and III. Aside from this difference, the other seven chapters across three textbooks are largely consistent (Fig. 3) and are regarded as the secondary knowledge subfields (F1-F7 in Table 4). Furthermore, the three textbooks contain 41, 47, and 31 sections in all chapters, and most of these sections are similar (Table S1). These sections are reviewed by the five experts one by one to identify 36 knowledge fields (Table 4). For example, the three textbooks include similar sections such as “A6-3 Contract Pricing Method”, “B7-3 Construction Project Contract Pricing Method”, and “C6-4 Construction Project Contract Pricing Method”, which collectively lead to the “F6-3 Contract Pricing Method” (Table 4). The specific correspondence between each section in the three textbooks and the 36 third-level knowledge fields is presented in Table S1. Consequently, a three-level CPM knowledge framework is obtained,

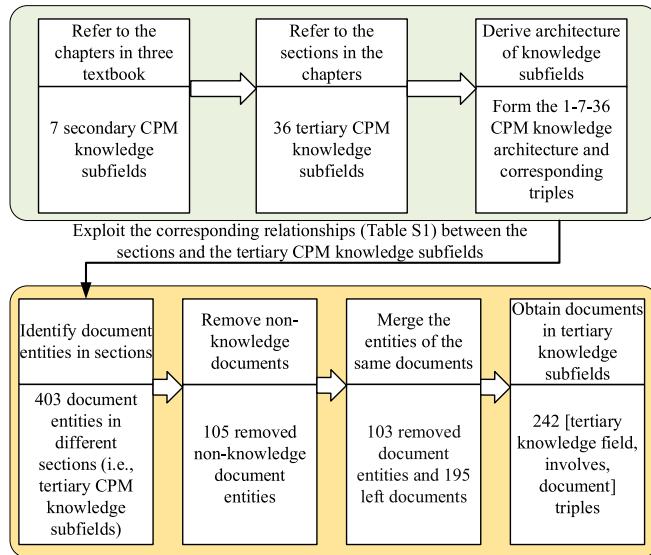


Fig. 3. Process of determining the subfields and documents of CPM.

comprising 1 primary knowledge field along with 7 secondary fields and 36 tertiary fields (Fig. 3). These 44 CPM knowledge fields constitute 7 [primary knowledge field, has subfield of, secondary knowledge field] triples and 36 [secondary knowledge field, has subfield of, tertiary knowledge field] triples (Fig. 2b).

The second step is to identify documents under each tertiary knowledge field using entity recognition and manual filtration (Fig. 3). Initially, a total of 403 document entities related to CPM standards, regulations, and laws are identified using the Chinese title symbols (i.e., 《》) [54,55,56]. Then, 105 of 403 documents (e.g., 《Contract of A Certain Project》) that do not refer to a specific CPM knowledge-related document are ruled out. Subsequently, the entities that refer to the same document are merged, which eliminates 103 document entities. For instance, the entity “Tendering and Bidding Law” is merged with “Tendering and Bidding Law of the People’s Republic of China”. Finally, 195 CPM-related documents are identified and coded (Table S1), such as “D004: Regulations on Quality Management of Construction Projects”. All document entities appear in specific chapters and sections in the textbooks, each corresponding to distinct subfields of CPM knowledge. As a result, a CPM-related document can easily be categorized into CPM knowledge subfields (Table 4). Notably, one CPM-related document may be categorized under multiple subfields. The 36 tertiary knowledge fields and 195 CPM-related documents form 242 [tertiary knowledge field, involves, document] triples (Fig. 2b and Table S2).

The third step is collecting, processing, and segmenting document content (Fig. 2b). The collection of specific content of 195 documents exploits the National Legal Database and government websites. 171 of 195 documents are purely textual, whereas the other 24 documents include 339 images. These images are uploaded to the cloud database (i.e., Amazon S3 in this research). A custom format, “</image> URL of the image in the cloud database <image>”, is created and inserted at the original location of the image within the text (Fig. 4). Following multimodal data processing, all document content is subdivided into chapters or sections to prevent the document content from becoming overly long. This procedure generates 1375 [document, contains, document content] triples (Fig. 2b). All details of the instance layer are in Table S2.

3.1.3. Pre-assessment, storage and sharing of the CPM-KG

The CPM-KG organizes 1,602 nodes and 1,696 links, containing 195 documents with 2.6 million words and 339 images (Fig. 2c). The CPM-KG is pre-assessed by a group of five CPM experts to ascertain whether there are significant omissions of CPM-related subfields or documents,

Table 4

The secondary and tertiary knowledge fields in CPM-KG.

7 secondary knowledge fields	36 tertiary knowledge fields	Number of field-related documents
F1: CPM fundamentals & Organization management in CPM	F1-1 Fundamentals of CPM F1-2 Organization of CPM F1-3 Planning and procurement modes of CPM F1-4 Development of CPM plans F1-5 Construction organization design F1-6 Dynamic control of CPM objectives F1-7 Construction project managers F1-8 Construction risks F1-9 Construction supervision	26 10 12 1 3 14 4 4 5
F2: Cost management in CPM	F2-1 Fundamentals of cost management F2-2 Cost planning F2-3 Cost control F2-4 Cost accounting F2-5 Cost analysis and evaluation	1 6 2 6
F3: Schedule management in CPM	F3-1 Fundamentals of schedule management and scheduling systems F3-2 Schedule objectives F3-3 Development and adjustment of schedules F3-4 Schedule control	8 11 4 4
F4: Quality management in CPM	F4-1 Fundamentals of quality management F4-2 Quality control F4-3 Project acceptance F4-4 Treatment of nonconformance F4-5 Mathematical and statistical methods in quality management F4-6 Quality supervision by governments	5 11 7 4 2 3
F5: Health, safety, and environmental management in CPM	F5-1 Fundamentals of occupational health and working environment F5-2 Safety management F5-3 Emergency management F5-4 Requirements of occupational health and working environment	3 15 5 5
F6: Contract management in CPM	F6-1 Fundamentals of contract, tendering, and bidding F6-2 Contract contents F6-3 Contract pricing methods F6-4 Contract risk, insurance, and guarantees F6-5 Execution of construction contracts F6-6 Dispute and claim management	5 18 2 10 6 6
F7: Information management in CPM	F7-1 Fundamentals of information management F7-2 Information classification, coding, and processing	6 4

Note: Fundamentals of each knowledge field refer to the definitions, objectives, procedures, tasks, or functions. More details of the instance layer are in Table S2.

and Table S2 shows the profiles of the experts. The CPM-KG is subsequently saved in the Neo4j-v5 database (Fig. 2c), which is the leading NoSQL graph database across both academic and industrial fields [57]. Besides storage, the CPM-KG has been shared in three ways: original Neo4j-formatted file, JSON-formatted dataset, and vectorized CPM-KG. This allows scholars and professionals to easily reuse the CPM-KG

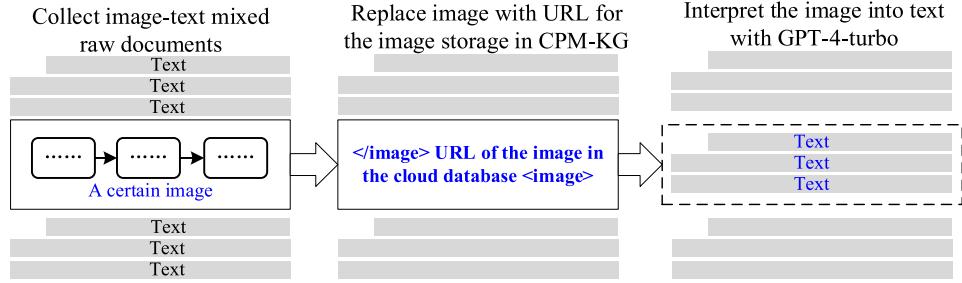


Fig. 4. Process of dealing with image-text mixed documents.

without needing to install the Neo4j-v5 database.

3.2. RAG-based integration of CPM-KG into GLMs

A three-stage integration process (Fig. 5) of GLMs and CPM-KG is devised with Langchain RAG framework [58]. The first of the three stages is creating the vectorized CPM knowledge repository (Fig. 5a).

Then, the question-relevant CPM knowledge will be found (Fig. 5b). Finally, it will formulate a prompt-knowledge-question query before feeding it into GLMs (Fig. 5c).

3.2.1. CPM-KG conversion to vectors

CPM-KG conversion to vectors starts with segmenting the CPM knowledge triples. The procedure begins by setting up a connection with

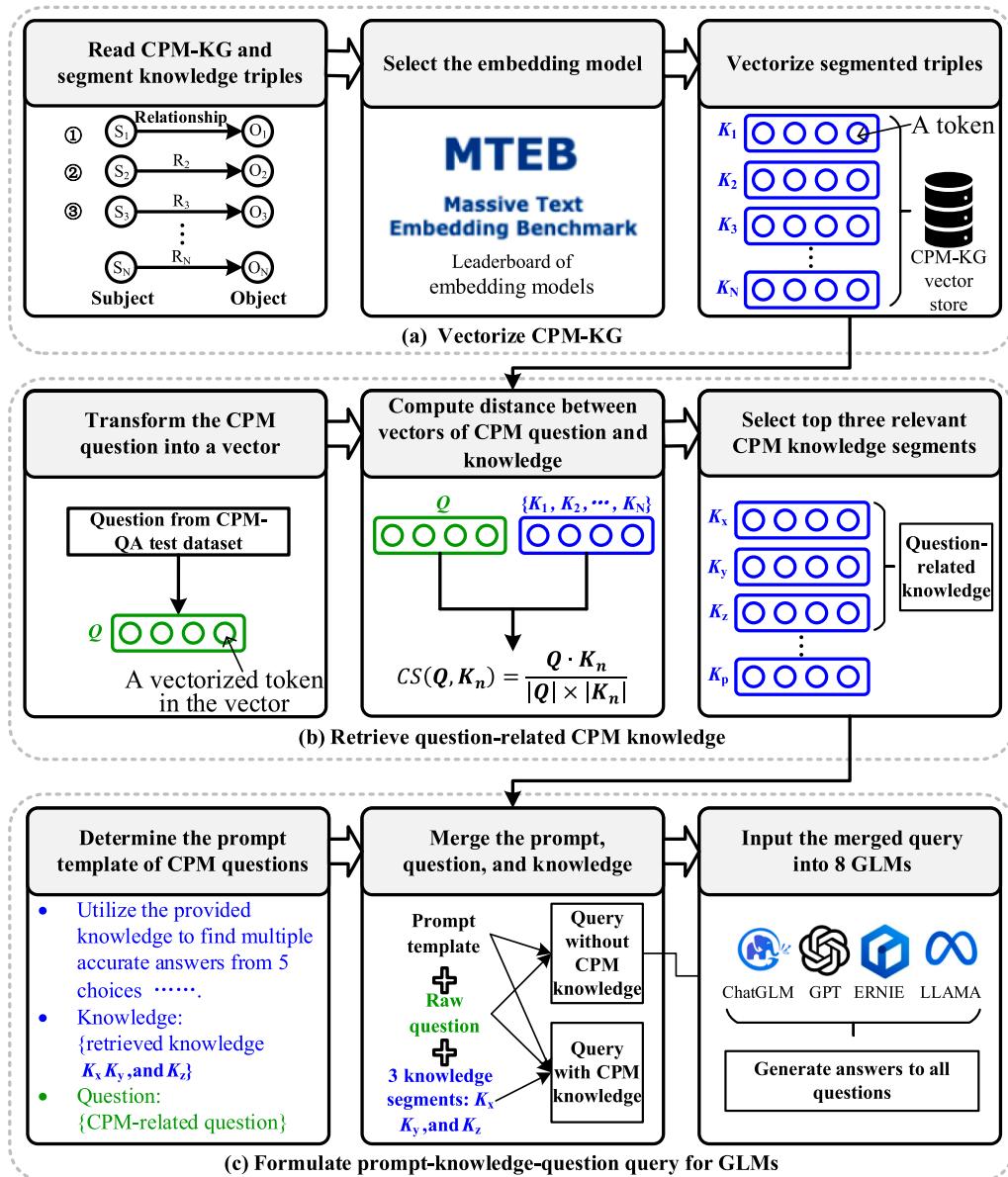


Fig. 5. RAG-based three-stage augmentation for GLMs using CPM-KG.

the multimodal CPM-KG through the Neo4j-v5 database platform. Next, knowledge triples are then extracted from the database and divided into segments, each containing 250 tokens with a 50-token overlap between adjacent segments [58,59].

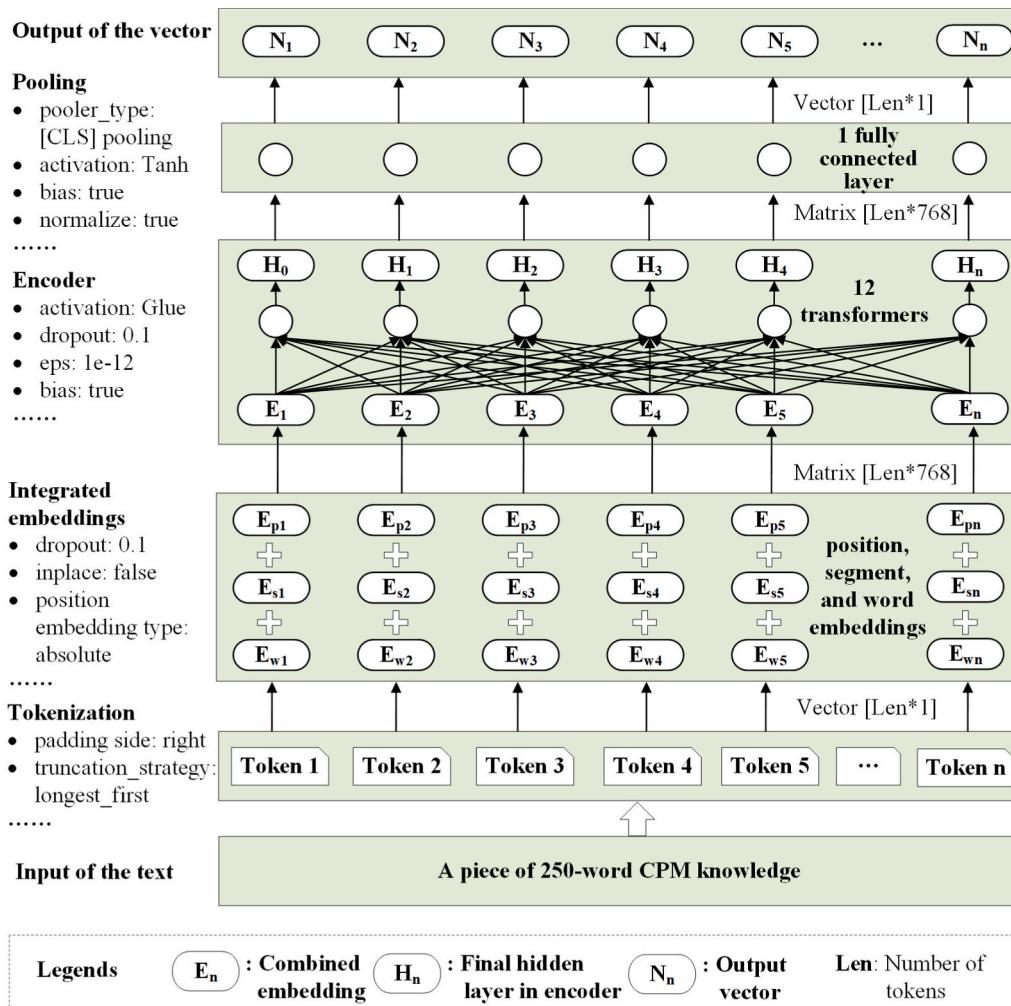
The selection of an embedding model primarily considers three factors: performance, maximum token length, and size [60]. Concerning the performance of an embedding model, existing model leaderboards can be referenced, which have comprehensive tests and are widely recognized by scholars worldwide [61]. The maximum token length refers to the longest input the embedding model can process [62]. If the maximum token length is limited, it may result in truncation of substantial portions of text, leading to information loss [63,64]. The model size should be as small as possible while ensuring optimal model performance to save computational resources and costs [65]. In this study, Dmeta-embedding-zh was selected as it ranked first among open-source embedding models on the Chinese leaderboard of the Massive Text Embedding Benchmark (MTEB) as of February 2024 [61], when the research was conducted. Furthermore, its maximum token length of 768 aligns with the study's requirements, and its compact model size of 110 million parameters makes it an efficient choice [66].

The vectorization process using the Dmeta-embedding-zh model involves inputting text, tokenization, combining embeddings, and pool-

ing, ultimately producing an output vector (Fig. 6). A 250-word piece of CPM knowledge is input first, which is tokenized using a pre-defined token vocabulary. Next, the word embeddings, sentence segment embeddings, and position embeddings are integrated. The integrated embedding then passes through 12 transformer layers (Fig. 6). Finally, the CLS-pooling strategy is applied, followed by normalization, resulting in the output vector \mathbf{K} . By repeating this process, the CPM-KG is transformed into a set of knowledge vectors $\{\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_N\}$ (Fig. 5a). Each step involves setting various hyperparameters, such as a dropout rate of 0.1, using GELU as the activation function in the encoder, and Tanh in the pooling layer. All parameter settings follow existing practices or retain the recommended default settings of the Dmeta-embedding-zh model [66,67]. More details on Dmeta-embedding-zh and its parameters are provided in the supplemental materials.

3.2.2. Question-relevant CPM knowledge searching

Retrieving the question-relevant CPM knowledge starts with transforming the CPM question into a vector (Fig. 5b). The question will be vectorized in the process in Fig. 6, with the same embedding model selected in Section 3.2.1. After that, the distances between the CPM question vector and knowledge segment vectors are computed, which help select the top three question-associated knowledge segments



Note: The exploited Dmeta-embedding-zh embedding model and the parameters are provided in the supplemental materials.

Fig. 6. The vectorization process from text to vector. Note: The exploited Dmeta-embedding-zh embedding model and the parameters are provided in the supplemental materials.

(Fig. 5b). Specifically, a question is converted into a question vector \mathbf{Q} using the embedding model corresponding to the tested GLM (Fig. 5b). The similarity between vectors \mathbf{Q} and \mathbf{K}_n is evaluated using widely recognized cosine similarity [68,69], defined as:

$$CS(\mathbf{Q}, \mathbf{K}_n) = \frac{\mathbf{Q} \bullet \mathbf{K}_n}{|\mathbf{Q}| \times |\mathbf{K}_n|} \quad (1)$$

When the top three question-associated knowledge segments are identified, they will have an effect in providing the context for answering the CPM question.

3.2.3. Formulate prompt-knowledge-question query for GLMs

This stage entails devising the prompt template, formulating the prompt-knowledge-question query, and submitting the merged query to GLMs (Fig. 5c). Considering the QA testing set includes questions that have either one or multiple accurate answers, two distinct prompt templates have been designed. The template for multi-answer questions is structured as: “Utilize the provided knowledge to find multiple accurate answers from 5 choices, strictly focusing on the choices without extra elaboration. Knowledge: {retrieved knowledge}. Question: {CPM-related question}.” In contrast, for questions requiring a single accurate answer, the template adjusts to “find the only one accurate answer from 4 choices”. Subsequently, the CPM knowledge and the initial question are inserted into the prompt template to generate a prompt-knowledge-question query (Fig. 5c). Lastly, this query is submitted to multiple GLMs (Fig. 5c), and Table 5 provides basic details of the eight up-to-date GLMs used in this research.

3.3. CPM-QA testing set construction

The development process of CPM-QA testing set (Fig. 1) involves (i) collecting CPM questions from CRCEEs, (ii) labeling the characteristics of each CPM question, and (iii) releasing the CPM-QA testing set.

3.3.1. CPM question collection

This study has developed a CPM-QA testing set to assess the performance differences between original GLMs and CPM-KG-augmented GLMs. The testing set consists of 2,435 multiple-choice questions from

Table 5
Basic information of 8 tested GLMs.

Tested GLM	Open source	Default embedding model during CPM-QA phase	Cutoff date of GLM development-used data	Parameter size	Seq. length
ChatGLM3-6B	✓	text2vec-large-chinese	2023.10	6B	8 K tokens
GPT-3.5-turbo	✗	ada-002	2022.01	20B	16 K tokens
GPT-4	✗	ada-002	2023.04	1.8 T*	32 K tokens
GPT-4-turbo	✗	text-embedding-3	2023.04	1.8 T*	128 K tokens
ERNIE-Bot	✗	Qianfan-Embedding	2023.03	20B	2 K tokens
ERNIE-Bot 4.0	✗	Qianfan-Embedding	2023.10	>1T*	10.2 K tokens
Qianfan-Chinese-Llama-2-7B	✗	Qianfan-Embedding	2023.07	7B	32 K tokens
Llama-2-70B-Chat	✓	Qianfan-Embedding	2023.07	70B	4 K tokens

Note: * means that the parameter scale is estimated by academia and industry, B: Billion, K: Kilo.

25 CRCEE papers spanning from 2013 to 2023. The CRCEE, a mandated authoritative qualification examination, is essential for qualifying as an engineer in the AEC industry. CRCEE encompasses Levels I and II, which are administered annually. Between 2020 and 2022, the outbreak of the coronavirus led to disruptions in exam scheduling across various provinces and cities, resulting in multiple CRCEEs occurring in the same year.

3.3.2. Annotations of CPM questions' characteristics

The CPM-QA testing set questions are manually tagged with four characteristics (Fig. 7). The characteristics include the CRCEE's level and year (Fig. 7a) and the related subfields of CPM knowledge (Fig. 7b). Additionally, a multiple-choice question (MCQ) in CRCEE has a single answer (SA) or multiple answers (MA), so these questions are divided into SAMCQ and MAMCQ (Fig. 7c). The examples of SAMCQ and MAMCQ are shown in Fig. 8. Furthermore, the presence of images within the questions is labeled, as indicated in Fig. 7d. These annotations (Fig. 7) can split the CPM-QA testing set into different question groups, thus offering a multifaceted view of GLMs' CPM-QA capabilities.

3.3.3. Release of the CPM-QA testing set

Given the shortage of CPM-QA-related testing sets, this valuable CPM-QA testing set (Table S3) is released on HuggingFace (Fig. 8), which is the globally foremost community for enthusiasts and researchers interested in GLMs. The CPM-QA testing set is shared in two versatile formats (CSV and JSON) to facilitate accessibility and encourage widespread use. Additionally, the comprehensive instructions for reutilizing the CPM-QA testing set have been meticulously outlined, ensuring users can fully leverage this testing set.

3.4. CPM-QA performance assessment and comparison of GLMs

The GLM performance investigation is split into three steps (Fig. 1). They are (i) formulating grading criteria for individual questions, (ii) determining the GLM performance evaluation indicator, and (iii) contrasting the CPM-QA performance differences of the original GLMs and CPM-KG-augmented GLMs.

3.4.1. Grading criteria of individual questions

This study implements the CRCEE grading method for both SAMCQs and MAMCQs. A SAMCQ is assigned a value of 1 mark, with s^s measured as follows:

$$s^s = \begin{cases} 1, & \text{if } A = T \\ 0, & \text{if } A \cap F \neq \emptyset \end{cases} \quad (2)$$

where A , T , and F imply the sets of answers from GLMs, accurate answers, and wrong answers. Additionally, a MAMCQ is assigned a value of 2 marks. The MAMCQ's score s^m is defined as:

$$s^m = \begin{cases} 2, & \text{if } A = T \\ 2 - 0.5 \times [n(T) - n(A)], & \text{if } A \subseteq T \cup A \neq T \\ 0, & \text{if } A \cap F \neq \emptyset \end{cases} \quad (3)$$

where n is the number of elements in corresponding sets.

3.4.2. Evaluation indicator of GLM performance

After calculating the scores of individual questions, the accuracy rate (AR), precision, recall, and F1-score are introduced to quantitatively assess the CPM-QA performance of GLMs on a designated question group. The study adopts four types of CPM question groups. They are the twenty-five groups comprising questions from different CRCEEs (Fig. 7a), thirty-six groups consisting of questions from diverse CPM subfields (Fig. 7b), two groups of SAMCQs and MAMCQs (Fig. 7c), and two groups of text-only and image-embedded questions (Fig. 7d). AR is the ratio of the total marks secured by the GLMs to the maximum marks achievable for a question group:

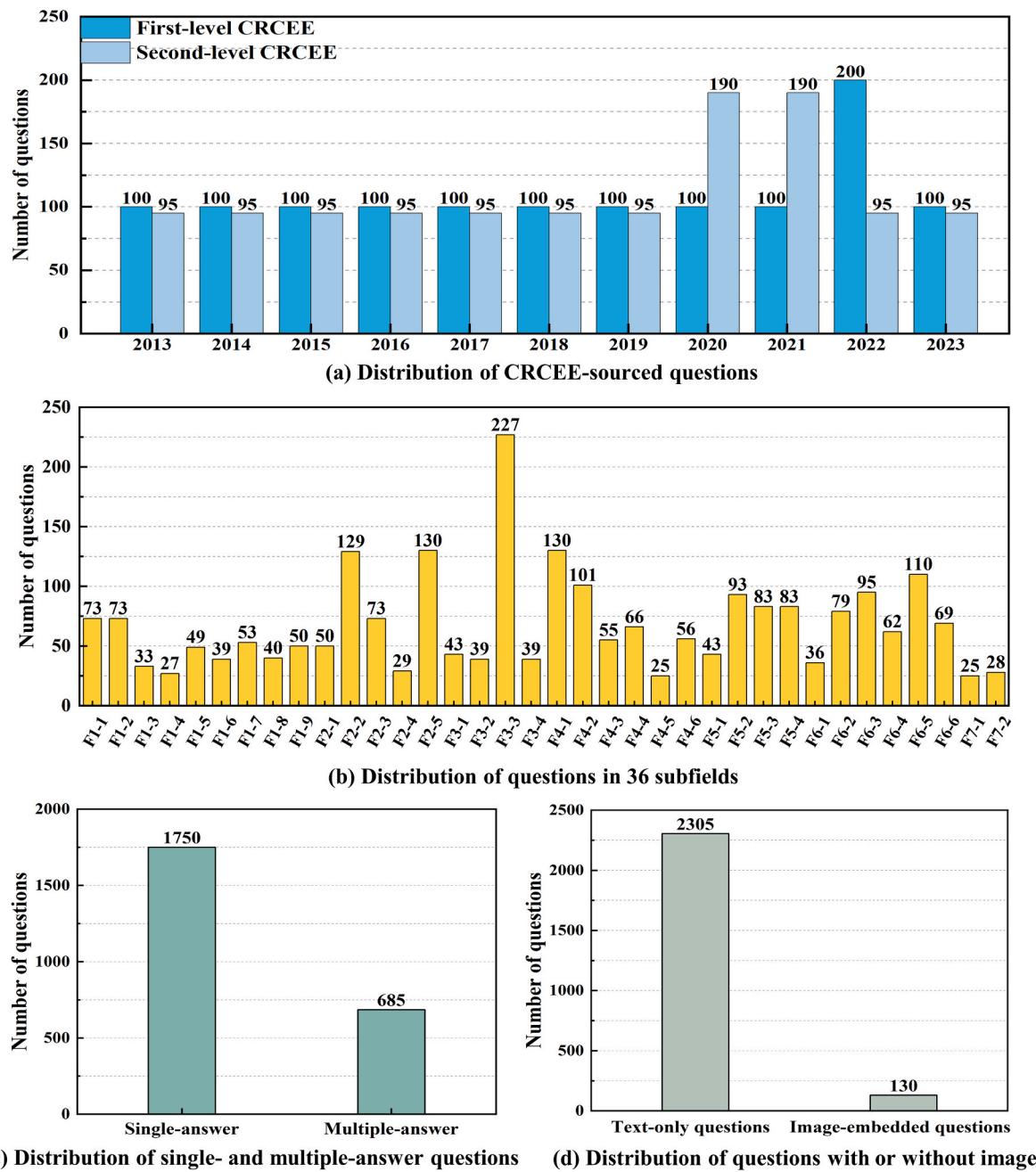


Fig. 7. Details of the 2,435-question CPM-QA testing set.

$$AR = \frac{\sum_{i=1}^{N^s} s_i^s + \sum_{j=1}^{N^m} s_j^m}{N^s \times 1 + N^m \times 2} \quad (4)$$

where s_i^s is the score of the i^{th} SAMCQ, s_j^m is the score of the j^{th} MAMCQ, and N^s and N^m represent the number of SAMCQs and MAMCQs in a CPM question group. Precision is the ratio of correct answers output by GLMs to the total number of answers produced. Recall is the ratio of correct answers output by GLMs to the total number of actual correct answers. The F1-score is the harmonic mean of precision and recall. They are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

where TP is true positive, FP is false positive, and FN is false negative.

3.4.3. CPM-QA performance comparison

To evaluate the variations in CPM-QA performance between original GLMs and CPM-KG-augmented GLMs, this work employs the Wilcoxon signed-rank test. This statistical approach is ideal for small samples and does not demand the assumption of data normality [70]. It is hypothesized that the CPM-QA performance of GLMs with CPM-KG does not significantly differ from those without CPM-KG, as opposed to the alternative hypothesis of a notable difference. SPSS is utilized as the tool for conducting the Wilcoxon signed-rank test.

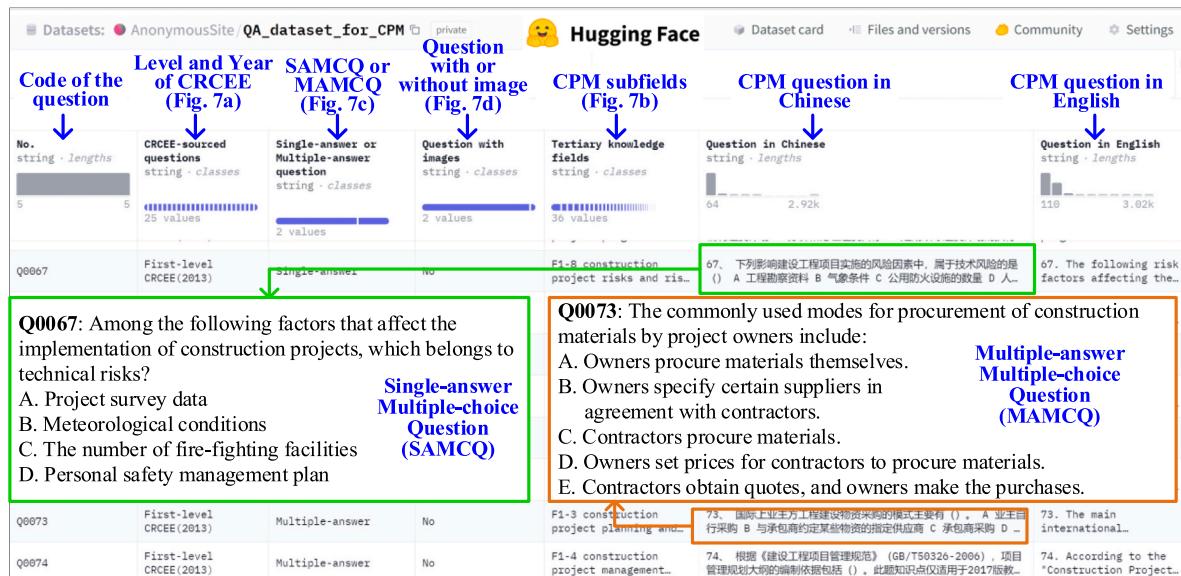


Fig. 8. Preview of the CPM-QA testing set.

4. Results

This section presents the results from three aspects: (i) demonstrating the improvement of AR, precision, recall, and F1-score of each GLM before and after the integration of multi-modal CPM-KG through 4 question groups (Fig. 7), (ii) comparing the performance of GLMs integrated with single- or multi-modal CPM-KG for CPM-QA, and (iii) comparing the performance of GLMs in case-based questions. After running the automatic QA as shown in Videos S1 and S2, the answers to multiple-choice questions and case-based questions from 8 GLMs are provided in Table S4.

4.1. CPM-KG-enabled GLM performance enhancements in CRCEEs

In all 25 CRCEEs during 2013–2023, the CPM-KG lifts the CPM-QA performance for all GLMs significantly. After CPM-KG injection, the AR of GLMs increases by 30.0 % on average, with variabilities from 13.3 % to 52.9 % (Fig. 9). Amongst the eight tested GLMs (Fig. 9), the three GLMs (ERNIE-Bot 4.0, GPT-4-turbo, and GPT-4) stand out after the CPM-KG empowerment, achieving average ARs of 0.682 (Fig. 9f), 0.643 (Fig. 9d), and 0.628 (Fig. 9c), surpassing the CRCEE's pass threshold of 0.600. The details of Wilcoxon signed-rank test of ARs' comparison are presented in Table S4. In addition, the precision, recall, and F1-score of GLMs have improved by 16.8 %, 16.1 %, and 16.4 %, respectively (Table 6). Among the GLMs, ERNIE-Bot 4.0 achieves the best precision, recall, and F1-score.

4.2. CPM-KG-enabled GLM performance enhancements across CPM subfields

CPM-KG notably lifts the performance of 8 GLMs responding to questions across 36 CPM subfields. From the viewpoint of GLMs, ERNIE-Bot 4.0 (Fig. 10f), GPT-4-turbo (Fig. 10d), and GPT-4 (Fig. 10c) perform the best on subfield-specific CPM questions, while Qianfan-Chinese-Llama-2-7B (Fig. 10g) holds the highest AR improvement of 56.3 %. From the viewpoints of 36 CPM subfields, CPM-KG brings an average AR improvement of 30.0 %, and the enhancements for these 36 subfields vary from 12.2 % to 57.8 % (Table S4). If GLM performance in a specific field does not improve, Fig. 10 can help indicate adding subfield-specific documents to the CPM-KG. The ARs of CPM-QA among 36 CPM subfields are different, strongly correlating with the proportion of MAMCQs in each subfield (Table 7). As MAMCQs are more challenging, the subfields

(e.g., F3-3) with a higher proportion of MAMCQs tend to exhibit low ARs (Table 7).

Across the 36 knowledge subfields, the CPM-QA improvements in precision, recall, and F1-score also demonstrate the effectiveness of the CPM-KG. The average improvement in precision, recall, and F1-score for the 8 GLMs is 18.6 %, 17.3 %, and 17.9 %, respectively (Table 8). Notably, the average precision, recall, and F1-score of ERNIE-Bot 4.0 are over 0.800 across 36 knowledge subfields, indicating that the CPM-KG effectively enhances answer precision while recalling most of the correct responses.

4.3. CPM-KG-enabled GLM performance enhancements on diverse question types

This study evaluates the CPM-QA performance enhancements of GLMs on SAMCQs and MAMCQs (Fig. 11), considering their varying difficulty levels. The combination of CPM-KG and GLMs facilitates an AR elevation for SAMCQs from 0.507 to 0.606, marking a 19.6 % improvement on average (Table S4). Additionally, ARs for MAMCQs rise from 0.267 to 0.395, reflecting a 48.0 % increase (Table S4). As the calculation of precision and recall for SAMCQs is consistent with that for the AR, Table 9 presents only the precision, recall, and F1-score values for MAMCQs. The average improvements in precision, recall, and F1-score for the eight GLMs are 8.8 %, 4.6 %, and 6.6 %, respectively (Table 9).

The combination of CPM-KG and GLMs exhibits marked performance advancements on both text-only and image-embedded questions. As shown in Fig. 12, Table 10, and Table 11, the average improvements for text-only questions are as follows: ARs increased by 30.6 %, precision by 17.8 %, recall by 16.2 %, and F1-score by 17.0 %. For questions containing images, the average improvements are as follows: ARs increased by 20.4 %, precision by 20.7 %, recall by 22.3 %, and F1-score by 21.4 %. Notably, except for GPT-4-turbo, the APIs of all other GLMs (e.g., Llama-2-70B-Chat) currently do not support image processing. This suggests that these indicators for image-embedded questions should actually be zero. The indicators presented in Fig. 12b are calculated using questions where the images have been transformed into text via GPT-4-turbo. This transformation mirrors the image-to-text process used in CPM-KG development (Fig. 4), enabling non-image-processing GLMs to handle image-embedded questions (Fig. 12b).

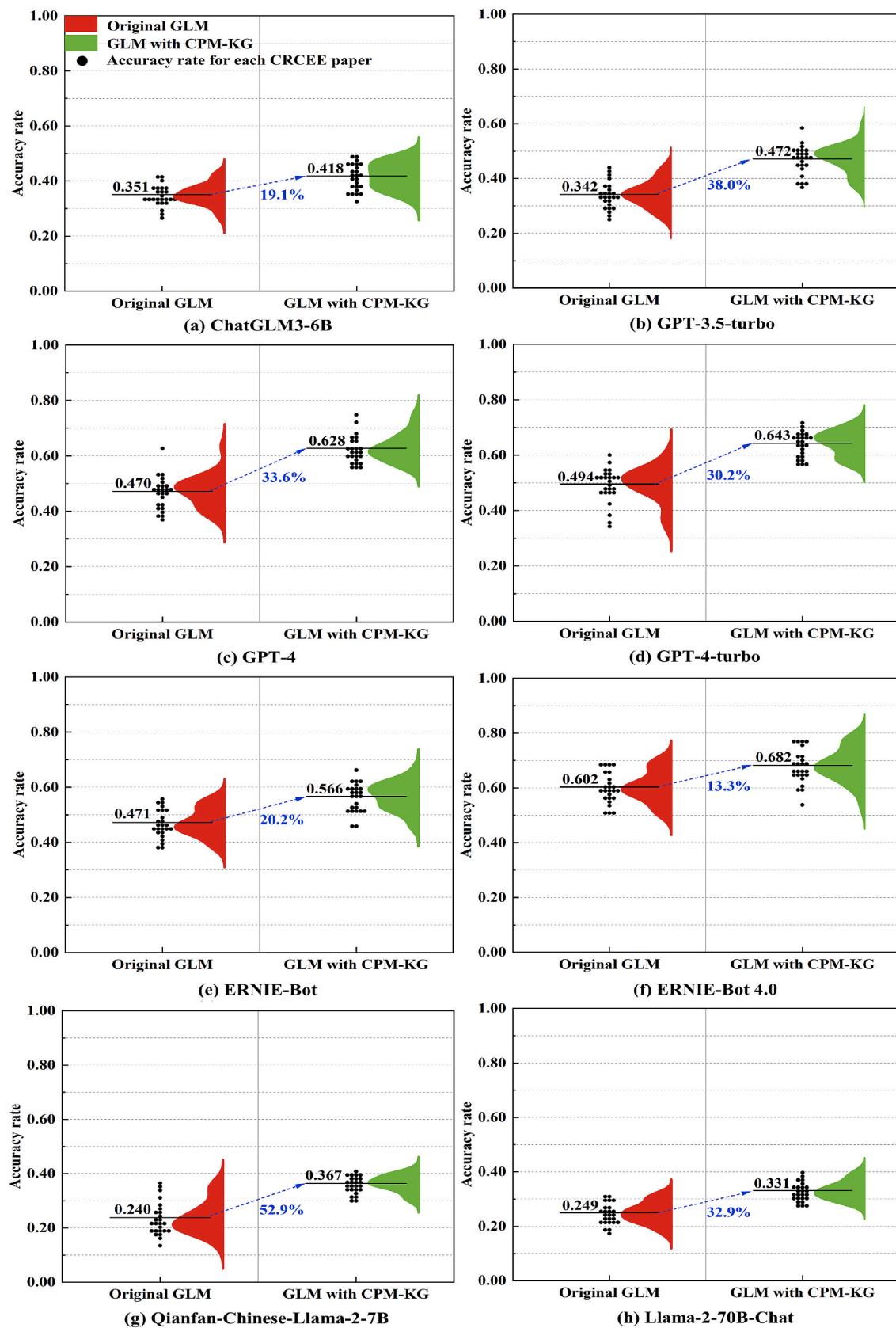


Fig. 9. AR enhancements of 8 GLMs with CPM-KG in 25 CRCEEs.

Table 6

Precision, recall, and F1-score of 8 GLMs in 25 CRCEEs.

Tested GLM	Precision		Recall		F1-score	
	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG
ChatGLM3-6B	0.507	0.566	0.468	0.536	0.487	0.551
GPT-3.5-turbo	0.503	0.606	0.486	0.569	0.494	0.587
GPT-4	0.635	0.725	0.654	0.724	0.644	0.724
GPT-4-turbo	0.643	0.754	0.654	0.750	0.649	0.752
ERNIE-Bot	0.673	0.734	0.714	0.774	0.693	0.753
ERNIE-Bot 4.0	0.745	0.788	0.733	0.798	0.739	0.793
Qianfan-Chinese-Llama-2-7B	0.389	0.536	0.373	0.496	0.381	0.515
Llama-2-70B-Chat	0.420	0.496	0.417	0.505	0.419	0.501
Average improvement amplitude	16.8 %		16.1 %		16.4 %	

4.4. GLM performance comparison between single- and multi-modal CPM-KGs

The single- and multi-modal CPM-KGs refer to the knowledge graphs without and with image data. The comparison results demonstrate that the inclusion of image data in the CPM-KG significantly enhances CPM-QA performance (Figs. 13 and 14). In all 25 CRCEEs during 2013–2023, multi-modal CPM-KG-based GLMs achieved an average AR improvement of 3.7 %, ranging from 1.9 % to 4.7 % (Fig. 13). Across 36 knowledge subfields, the AR enhancement achieved by multi-modal CPM-KG-based GLMs is 3.9 % (Fig. 14). Given that a large portion of knowledge remains in textual form, the overall performance improvement due to image-based knowledge is moderate. The statistical analysis using Wilcoxon signed-rank test confirms that these performance improvements are significant (Table S4). Notably, the subfields with a larger number of figures show a more significant performance enhancement. For example, the AR enhancement of F1-4 Development of CPM plans (Table 4) has increased by 13.5 %. The details of Wilcoxon signed-rank test of ARs' comparison are presented in Table S4.

4.5. CPM-KG-enabled GLM performance enhancements in case-based questions

Besides the 2,435 multiple-choice questions, this study also extracts 100 case-based questions (Table S3) from the “Professional Engineering Management and Practice” subject in CRCEEs. These 100 questions are open-ended and more representative of real-world CPM scenarios. The scores, grading methods, and standard answers of these questions are consistent with those used in the official examinations. Two experts (Table S4) independently score the answers, and the average score is taken. If the difference between the two experts' scores exceeds one point, a third expert (Table S4) joins them to discuss and determine the final score.

The results show that CPM-KG helps GLMs better respond to open-ended case-based questions (Fig. 15). In the 100 case-based questions, multi-modal CPM-KG-based GLMs show an average AR improvement of 11.2 %. Among the GLMs, ERNIE-Bot 4.0 achieves the highest AR of 0.822, while GPT-3.5-turbo shows the greatest improvement, with a 26.0 % increase. Five out of eight GLMs with CPM-KG surpass the 0.600 threshold. These results suggest that CPM-KG has the potential to help GLMs deal with practical engineering problems to a certain extent. Table S4 provides the details of the Wilcoxon signed-rank test.

5. Discussion

5.1. Critical contributions

This CPM-QA research has made critical contributions in the following three aspects: (i) revealing the significance of a CPM-KG in augmenting CPM-QA capabilities of GLMs, (ii) developing a reusable multimodal CPM-KG that can lift the CPM-QA proficiency across GLMs, and (iii) creating a CPM-QA testing set that serves as a benchmark.

This research reveals the significance of CPM-specific knowledge in enhancing GLM-based CPM-QA. Prior research on CPM-QA is constrained by grammar rules and conventional machine-learning models for narrow-scoped QA tasks (Table 1), failing to address complex CPM-QA involving dozens of subfields (Fig. 2). The combination of CPM-KG and prominent GLMs (i.e., ERNIE-Bot 4.0, GPT-4-turbo, and GPT-4) attains superhuman performance in CPM-QA (Fig. 9), surpassing CRCEEs with a mere 7 % human success rate [71]. Although the remaining five GLMs do not meet the criteria for passing the examinations, their performance also has significant improvements (Fig. 9).

Moreover, a multimodal CPM-KG (Fig. 2) is created to be reused for any GLM-based CPM-QA applications. Existing research (Table 1) often develops knowledge repositories of a smaller scale, which usually feature fewer than ten isolated documents. These knowledge repositories also consider only text information, overlooking image data (Table 1). In contrast, our multimodal CPM-KG employs a concise conceptual layer to organize 36 tertiary knowledge fields, 195 CPM documents, over 2.6 million words, and 339 images (Fig. 2). The efficacy (Fig. 9) and comprehensiveness (Fig. 10) of CPM-KG have been statistically validated. Notably, the CPM-KG is provided in Neo4j, JSON, and Chroma formats within the supplemental materials, making it easily reusable to empower more GLMs for CPM-QA of China beyond the GLMs in Table 5.

This work finally contributes to offering a CPM-QA testing set (Fig. 7), consisting of 2,435 CPM-related multiple-choice questions. Previous research developed task-specific, non-public QA testing sets (Table 1), covering far fewer than 36 CPM subfields in Table 4. The shortage of QA testing sets makes many GLM performance evaluation works have to rely on the subjective Elo rating approach [72,73]. This may encounter significant and sometimes contradictory variations in GLM performance assessment results among different experts [74]. This study not only builds a CPM-QA testing set based on CRCEEs but also manually annotates the characteristics of each question (Fig. 7). This testing set can be used by subsequent researchers as a benchmark for CPM-QA studies, helping to propel advancements in the field of intelligent CPM-QA (Fig. 7).

5.2. Updatability of CPM-KG

Due to the continuous development of CPM knowledge, it is difficult to build a perfect CPM-KG. Updating CPM-KG can reduce this limitation of CPM-KG. The updating of multimodal CPM-KG includes two aspects: (i) adjusting the three-level CPM knowledge architecture and (ii) modifying the nodes of CPM documents.

The three-level knowledge architecture (Table 4) within CPM-KG is designed with flexibility, enabling it to be adjusted to accommodate varying needs across different contexts. For example, as the construction industry grows and diversifies, emerging subfields (e.g., F7-2 Information classification, coding, and processing) can be incorporated into the knowledge architecture. In regions outside of China, where local construction practices and regulations may differ, the knowledge architecture can be tailored by adding or modifying subfields that are more

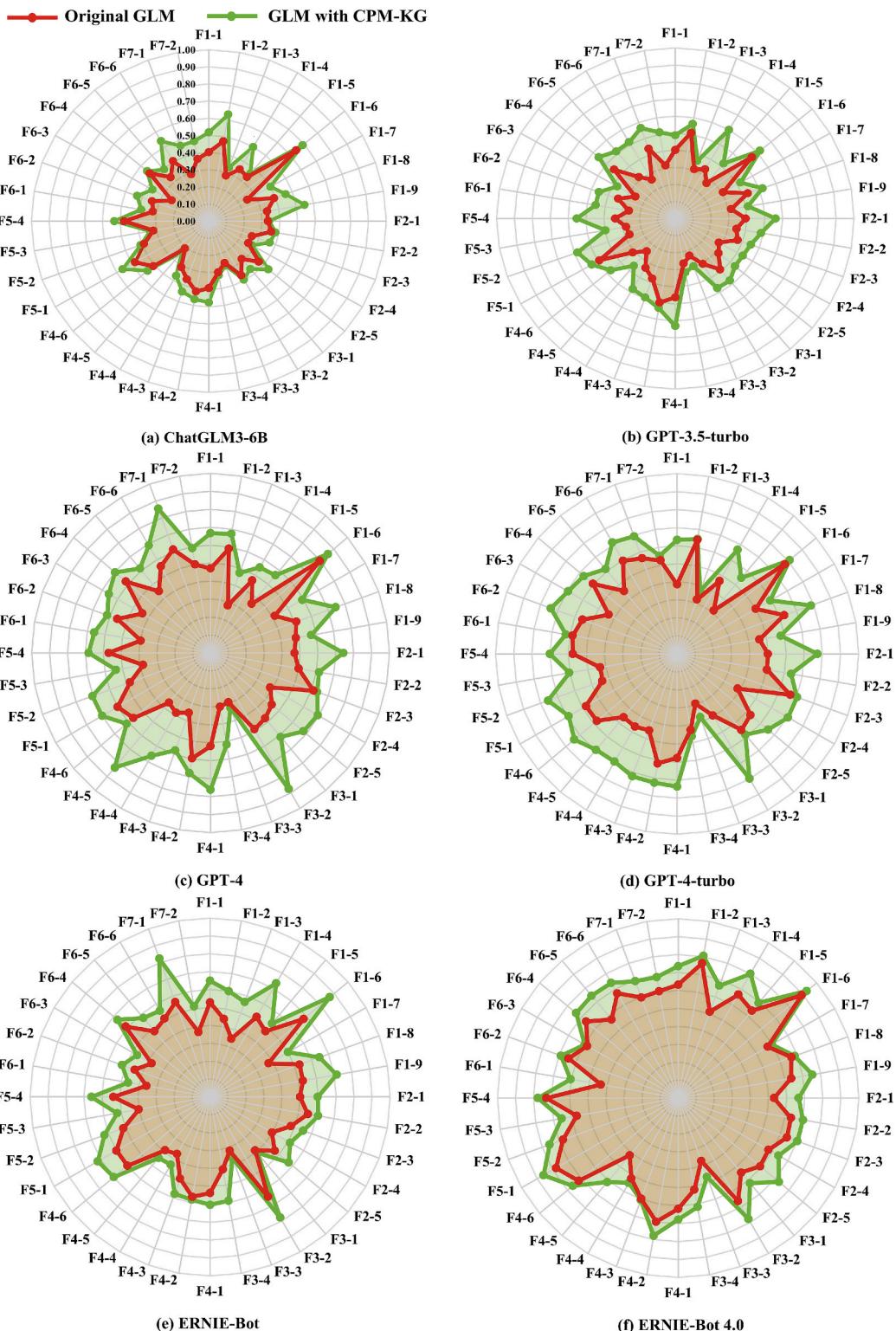


Fig. 10. AR enhancements of 8 GLMs with CPM-KG across 36 CPM subfields.

pertinent to the local context. This updatability ensures that the CPM-KG incorporates new and regionally relevant CPM subfields.

Besides adjusting the three-level knowledge subfields, the document nodes within the CPM-KG (Fig. 2) can also be modified to reflect changes in CPM. For example, the standards, regulations, and laws are often subject to revisions and updates. Once these changes occur, the corresponding document nodes in the CPM-KG (Fig. 2) can be adjusted

accordingly. If new documents (e.g., standards, regulations, or laws) are issued, they can also be integrated into the CPM-KG provided in the supplemental material. This updatability allows CPM-KG to stay up-to-date and adapt to changes in the construction industry.

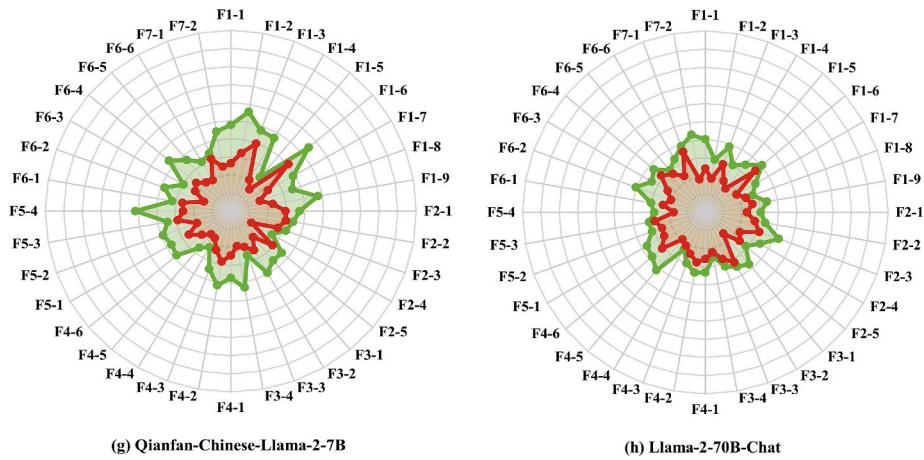


Fig. 10. (continued).

Table 7
Correlation between the MAMCQs and the ARs of GLMs.

Indicators		ARs of original GLMs	ARs of GLMs with CPM-KG
Proportion of MAMCQs	Pearson Correlation	-0.665 ^{**}	-0.572 ^{**}
	Sig. (2-tailed)	0.000	0.000
	Sum of Squares and Cross-products	-0.174	-0.146
	Covariance	-0.005	-0.004
	Number of subfields	36	36

Note: ** implies that correlation is significant at the 0.01 level.

5.3. Prototype deployment and preliminary validation

A prototype based on the proposed approach has been developed and deployed. Functional validation of the prototype is conducted ([Table S5](#)). Three experienced testers have been invited to evaluate the functionality of the 13 features of the prototype, including switching dialogue modes (i.e., QA based on GLMs with or without an external knowledge base), selecting GLMs, adjusting the temperature parameter, configuring historical dialogue rounds, and others. The testers are given access to the complete test set, allowing them to experiment with each feature individually. Following this, the testers confirm the functionality of all features ([Table S5](#)). A concrete example of QA based on GLM with CPM-KG is provided in [Fig. 16](#), where the user inputs a question, and the GLM outputs the corresponding answer. Additionally, the three pieces of knowledge retrieved during the RAG process are displayed in [Fig. 16](#). The deployment code and demonstration video of the CPM-QA prototype are included in the [supplementary materials](#).

The smart CPM-QA system in this study has the potential to serve as a reference for engineers in addressing four types of questions: (i) law and regulation queries, (ii) compliance checking, (iii) case analysis, and (iv)

operational guidance consultations. For law and regulation queries, such a smart QA system allows engineers to search for specific laws, regulations, or standards (Fig. 17a). For compliance checking, it helps assess whether project practices or components comply with relevant regulatory requirements (Fig. 17b). Concerning case analysis, the smart QA approach enables users to analyze specific scenarios and receive informed responses (Fig. 17c). Finally, for operational guidance, it provides practical steps and procedures for project management implementation (Fig. 17d). As current smart QA systems are not flawless, the prototype reminds users that the provided responses are for reference purposes only.

5.4. Measures for enhancing explainability and reducing bias

The CPM-KG-based RAG in this study can enhance the explainability of GLM results. RAG has been demonstrated to be an effective method for increasing the transparency of responses [36]. As CPM-KG provides three pieces of relevant knowledge to GLMs, users can verify whether or not a response is grounded in the retrieved knowledge. For instance, the GLM's answer (i.e., "Request the provision of documents and materials related to project quality ...") in Fig. 16 is derived from the second piece of retrieved knowledge, thereby clarifying the origin of this answer. In this study, the three retrieved knowledge segments for each question are presented in Table S4, contributing to enhanced explainability.

Furthermore, the study introduces a self-explanation prompt to enhance the explainability of GLM responses. The self-explanation prompt is essentially the original prompt with the addition of “Please explain the reasons for the answer”. This enables GLMs to generate the answer, accompanied by an explanation of the reasoning behind their answer generation process. The self-explanation prompt provides deeper insights into the logic underlying the GLM’s output and helps users understand the rationale behind specific responses. Table S5 presents ten self-explanation samples. Using MCO-0323 (Fig. 17a) as an example,

Table 8
Precision, recall, and F1-score of 8 GLMs across 36 CPM subfields.

Tested GLM	Precision		Recall		F1-score	
	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG
ChatGLM3-6B	0.510	0.578	0.470	0.544	0.489	0.561
GPT-3.5-turbo	0.514	0.618	0.491	0.581	0.502	0.599
GPT-4	0.647	0.750	0.669	0.751	0.658	0.751
GPT-4-turbo	0.653	0.772	0.667	0.771	0.660	0.771
ERNIE-Bot	0.682	0.754	0.723	0.793	0.702	0.773
ERNIE-Bot 4.0	0.740	0.800	0.742	0.803	0.741	0.801
Qianfan-Chinese-Llama-2-7B	0.392	0.550	0.377	0.512	0.384	0.530
Llama-2-70B-Chat	0.415	0.506	0.416	0.511	0.415	0.509
Average improvement amplitude	18.6 %		17.3 %		17.9 %	

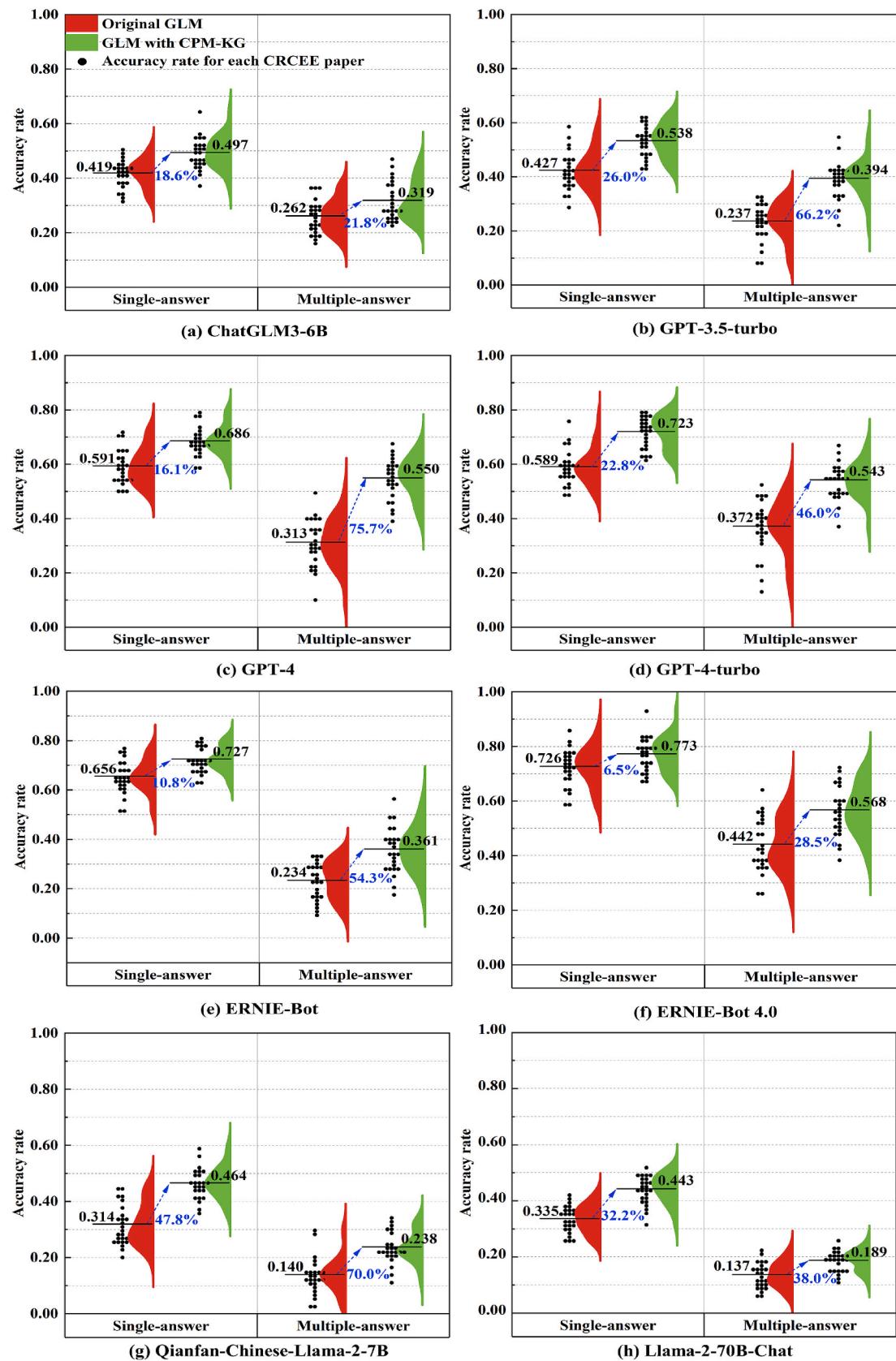


Fig. 11. AR enhancements of 8 GLMs with CPM-KG on SAMCQs and MAMCQs.

Table 9

Precision, recall, and F1-score of 8 GLMs on MAMCQs.

Tested GLM	Precision		Recall		F1-score	
	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG
ChatGLM3-6B	0.729	0.749	0.591	0.635	0.653	0.687
GPT-3.5-turbo	0.704	0.796	0.643	0.665	0.672	0.725
GPT-4	0.738	0.827	0.809	0.823	0.772	0.825
GPT-4-turbo	0.772	0.844	0.816	0.828	0.793	0.836
ERNIE-Bot	0.720	0.757	0.866	0.900	0.786	0.822
ERNIE-Bot 4.0	0.802	0.811	0.805	0.818	0.803	0.815
Qianfan-Chinese-Llama-2-7B	0.566	0.712	0.511	0.573	0.537	0.635
Llama-2-70B-Chat	0.630	0.637	0.629	0.659	0.630	0.648
Average improvement amplitude	8.8 %		4.6 %		6.6 %	

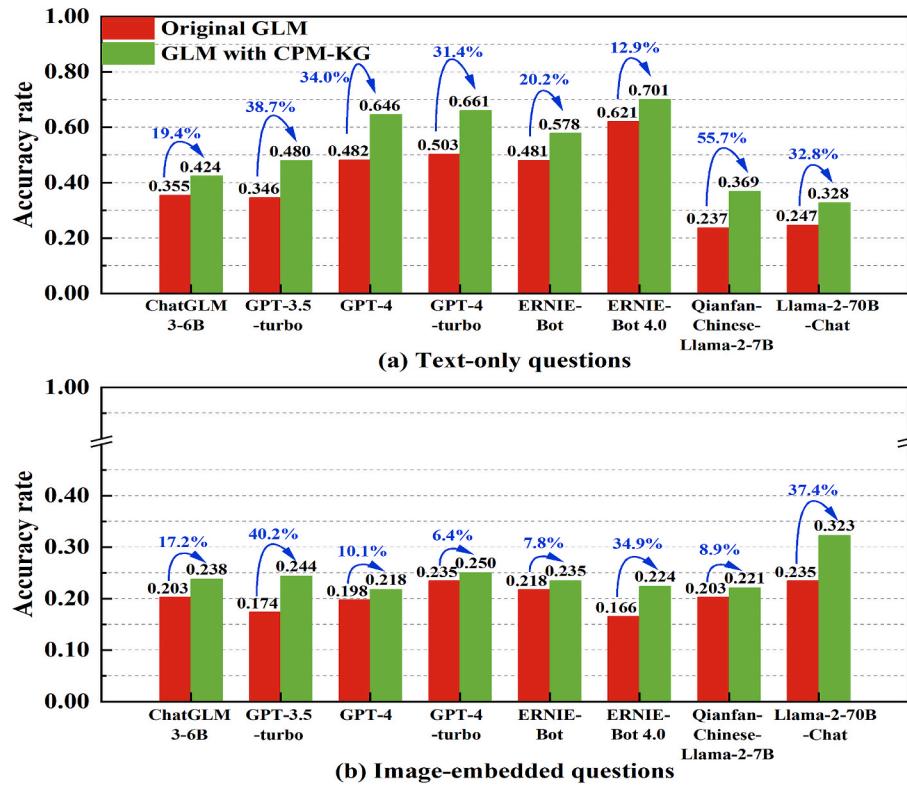


Fig. 12. AR enhancements of 8 GLMs with CPM-KG on text-only and image-embedded questions.

Table 10

Precision, recall, and F1-score of 8 GLMs on text-only questions.

Tested GLM	Precision		Recall		F1-score	
	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG
ChatGLM3-6B	0.514	0.576	0.474	0.542	0.653	0.687
GPT-3.5-turbo	0.512	0.619	0.494	0.581	0.672	0.725
GPT-4	0.650	0.745	0.671	0.745	0.772	0.825
GPT-4-turbo	0.655	0.774	0.669	0.769	0.793	0.836
ERNIE-Bot	0.687	0.751	0.726	0.789	0.786	0.822
ERNIE-Bot 4.0	0.754	0.801	0.754	0.803	0.803	0.815
Qianfan-Chinese-Llama-2-7B	0.390	0.544	0.376	0.507	0.537	0.635
Llama-2-70B-Chat	0.415	0.500	0.412	0.504	0.630	0.648
Average improvement amplitude	17.8 %		16.2 %		17.0 %	

each GLM explains how it refers to the “Bidding Law of the People’s Republic of China” to generate its answer (Table S5).

Beyond leveraging RAG and self-explanation to enhance performance with CPM-KG, three additional measures implemented in this study also help to mitigate potential concerns on model bias and decision-making opacity. These measures include (i) specially designed

CPM-QA prompts, (ii) updates to CPM-KG, and (iii) cross-referencing multiple answers from different GLMs. Specifically, the specially designed prompts in Section 3.2.3 instruct GLMs to rely on the provided background knowledge for their answer, minimizing hallucinations. Furthermore, as the CPM knowledge may become outdated as discussed in Section 5.2, CPM-KG in this study is designed to be updatable and

Table 11

Precision, recall, and F1-score of 8 GLMs on image-embedded questions.

Tested GLM	Precision		Recall		F1-score	
	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG	Original GLM	GLM withCPM-KG
ChatGLM3-6B	0.374	0.428	0.355	0.432	0.364	0.430
GPT-3.5-turbo	0.341	0.388	0.321	0.362	0.331	0.374
GPT-4	0.333	0.371	0.331	0.359	0.332	0.365
GPT-4-turbo	0.410	0.414	0.374	0.430	0.391	0.422
ERNIE-Bot	0.419	0.436	0.485	0.511	0.449	0.471
ERNIE-Bot 4.0	0.340	0.692	0.342	0.701	0.341	0.696
Qianfan-Chinese-Llama-2-7B	0.354	0.389	0.304	0.315	0.328	0.348
Llama-2-70B-Chat	0.462	0.499	0.497	0.530	0.479	0.514
Average improvement amplitude	20.7 %		22.3 %		21.4 %	

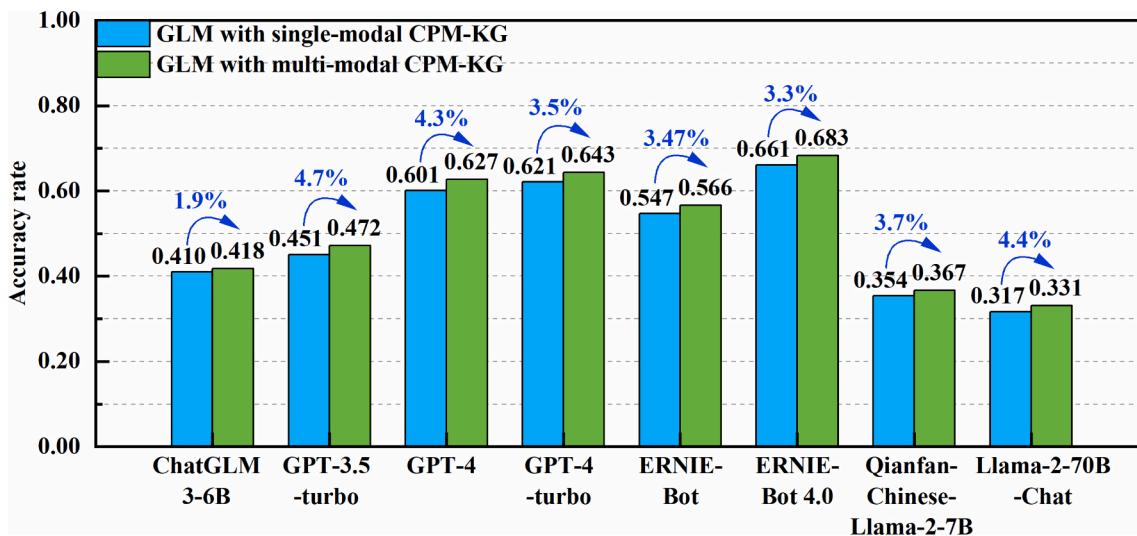


Fig. 13. AR enhancements of 8 GLMs on single- and multi-modal CPM-KGs in 25 CRCEEs.

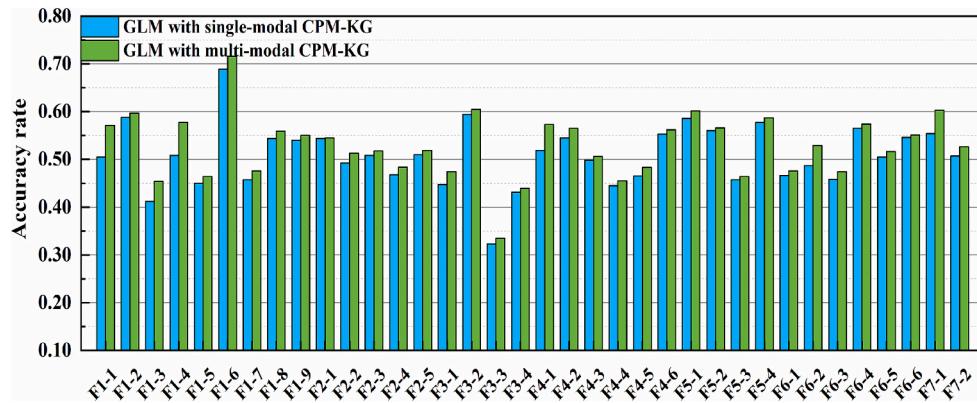


Fig. 14. AR enhancements of GLMs with single- and multi-modal CPM-KGs across 36 subfields.

adjustable to avoid biased answers resulting from outdated knowledge. Finally, this research utilizes multiple GLMs to answer the same question (Figs. 9–15). In practical applications, the prototype allows for switching between different GLMs, enabling users to compare responses from different GLMs. This comparative approach reduces the likelihood of errors associated with any single GLM.

6. Conclusion

This research augments GLMs on CPM-QA with CPM-KG. It involves (i) the development of a multimodal CPM-KG across 36 CPM subfields,

(ii) the orchestration of CPM-KG and GLMs in three stages, (iii) the construction of a 2,435-question CPM-QA testing set from CRCEEs, and (iv) evaluation and comparison of GLMs performance before and after CPM-KG integration. The CPM-KG enhances the GLM performance of CPM-QA by 30.0 % in all CRCEEs, enabling top-performing CPM-KG-augmented GLMs (i.e., ERNIE-Bot 4.0, GPT-4-turbo, and GPT-4) to pass the CRCEEs (Fig. 9). CPM-QA performance enhancements within 36 CPM subfields are between 12.2 % and 57.8 % (Fig. 10). For different question types, CPM-KG yields improvements of 19.6 % for SAMCQs, 48.0 % for MAMCQs, 30.6 % for text-only questions, and 20.4 % for image-embedded questions. The multimodal CPM-KG also outperforms

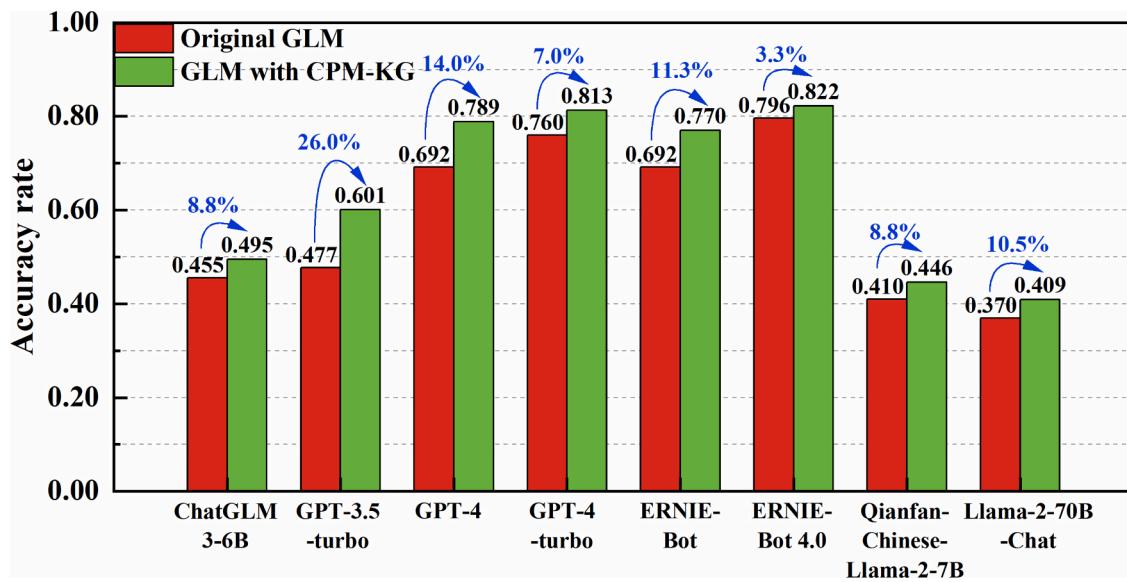


Fig. 15. AR enhancements of 8 GLMs on case-based questions.

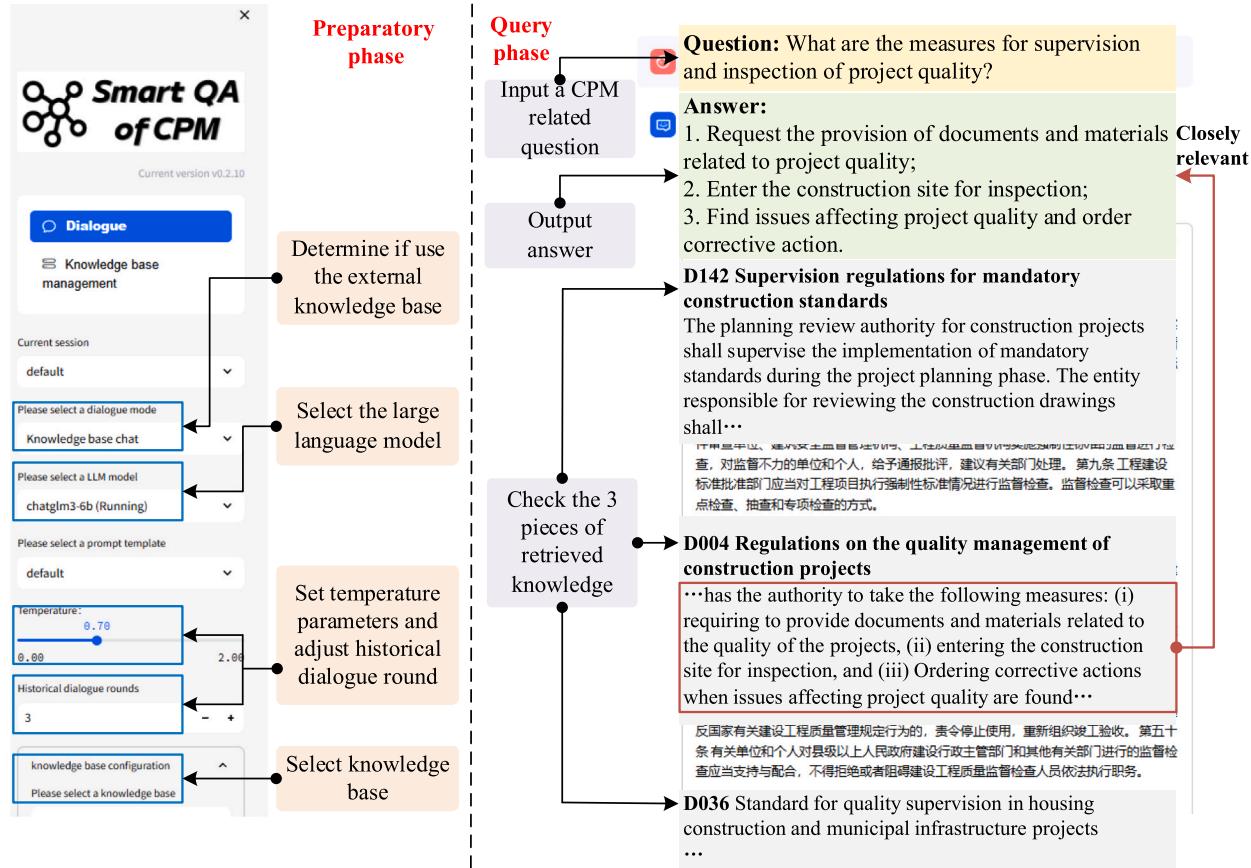


Fig. 16. A CPM-QA example in the prototype.

the text-only single-modal CPM-KG in enhancing CPM-QA performance. This study reveals the significance of CPM-KG in enhancing CPM-QA for GLMs, shares a plug-and-play CPM-KG, and delivers an openly available CPM-QA testing set.

This study has potential impacts on both academia and industry, primarily through (i) offering an additional CPM query channel, (ii) lowering the threshold for CPM queries, and (iii) facilitating the

development of the CPM-QA research field. Firstly, this research provides an additional CPM-QA channel beyond traditional approaches that rely on literature reviews, search engines, or expert consultations. Moreover, this channel is more efficient and cost-effective compared to traditional ways. Secondly, the GLM augmented by CPM-KG has passed professional engineering examinations and has demonstrated effectiveness in handling basic CPM queries (Figs. 9–15). The prototype could be

MCQ-0323: According to the "Bidding Law of the People's Republic of China" and related regulations, what are the requirements of the bidder for projects that require bidding?

根据《中华人民共和国招标投标法》及相关法规, 对必须招标的项目, 招标人行为有哪些具体要求?

Answer:

1. Follow the principles of openness, fairness, justice, and good faith.
2. Do not organize site visits for individual or selected potential bidders.
3. ...

8. 投标人不得就同一招标项目向潜在投标人提供有差别的项目信息。
 9. 投标人不得挪用投标保证金。
 10. 投标人可以自行决定是否编制标底。一个招标项目只能有一个标底。标底必须保密。
 11. 招标委托编制标底的中介机构不得参加委托编制标底项目的投标, 也不得为该项目的投标人编制标文件或者提供咨询。
 12. 投标人不得限定或者指定特定的品牌。

(a) Law and regulation query

CBQ-013: The owner issues the bidding document for a new project, with the following terms: Please identify any issues in the bidding document.

建设单位发布某新建工程招标文件, 部分条款有: 发包范围为土建、水电、通风空调、消防、智能化等工程。实行施工总承包管理; 投标限額为65000.00万元, 答约金額为1500.00万元, 工程款按进度完成工作的量%支付; 质量保證金为%; 答約保證金为15%; 钢材报庄采购本市钢厂的产品; 消防及通风空调专项工程合同金额1200.00万元, 由建设单位指定发包, 总承包服务费3.00%。请指出招标文件中的不妥之处, 并说明理由。

Answer:
 The issues in the bidding document are as follows:
 1. The scope in the bidding document lacks...
 2. The provisional amount is 15 million yuan...
 ...

(c) Case analysis

CBQ-012: The owner delays a payment of 10 million yuan for 30 days. The contractor claims 49,000 yuan in interest compensation based on the agreed rate. Please determine whether the above claim is valid.

因建设单位延迟30天支付进度款1000万元, 项目部按约定利率索赔利息4.9万元。请判断上述索赔事项是否成立。

Answer:
 If the owner fails to make the progress payment by the agreed deadline, the contractor shall promptly issue a notice to the owner requesting payment...
 其次, 需要计算出工程停工导致的直接损失, 包括现场管理人员工资、奖金和现场工人窝工补偿费。这些损失的计算需要依据相关的人工成本和窝工费用的标准, 进行详细的计算。
 最后, 需要将计算出的损失与工程合同中的相关条款进行对比, 确认这些索赔事项是否符合合同中的规定, 如果符合, 那么索赔事项就是成立的, 反之则不成立。

(b) Compliance checking

CBQ-011: What are the basis for estimating the project cost?

工程报价编制依据有哪些?

Answer:
 The basis for estimating the project cost includes the following:
 1. Bill of quantities:...
 2. Regulations of project pricing :...
 3. Information of marketing price ;...
 ...
 5. 类似项目的成本资料: 类似项目的成本资料可以为工程报价编制提供参考和启示。
 综上所述, 工程报价的编制依据包括工程量清单、工程计价有关规定、市场价格信息、相关定额和类似项目的成本资料等。

(d) Operational guidance consultation

available online and accessible 24/7, significantly lowering the entry barrier for CPM queries and thus facilitating the widespread dissemination of CPM knowledge. Lastly, the study provides a benchmark dataset for comparing the results of CPM-QA studies conducted by different researchers. Without such datasets, measuring and evaluating progress across various CPM-QA studies is difficult. This research fills this gap, promoting further advancement in the CPM-QA field.

Although this study demonstrates that CPM-KG significantly enhances the CPM-QA performance of GLMs, achieving fully intelligent CPM-QA remains a distant and ambitious goal. Future research will focus on (i) incorporating additional data modalities into CPM-KG, (ii) continuous integration with cutting-edge GLMs, (iii) improving the explainability of GLMs' answers, and (iv) validating the application efficiency of the developed prototype. Specifically, while CPM-KG in this research handles image data, further exploration into additional modalities such as video and audio would be valuable. Secondly, despite the promising improvements in CPM-QA (Figs. 9–15), integrating advanced GLMs like the forthcoming GPT 5.0 and ERNIE 5.0 is expected to further enhance CPM-QA performance. Thirdly, although both the CPM-KG-based RAG and self-explanation improve the explainability of GLMs' answers, the precise mechanisms and explanations of how the GLM generates such answers remain unresolved, requiring ongoing research efforts from global academia and industry. Finally, the application validation of the smart CPM-QA prototype consists of two main aspects: functional validation and efficiency validation. While functional validation has been conducted, the efficiency validation requires extensive practical data accumulated over time for comprehensive

testing, which is currently difficult to achieve and represents a future research direction.

7. Glossary

The abbreviations used in this study and their explanations are shown in Table 12.

Table 12
List of abbreviations.

Abbreviation	Explanation
AEC	Architecture, Engineering & Construction
AR	Accuracy Rate
CPM	Construction Project Management
CPM-QA	Construction Project Management Question-Answering
CPM-KG	Construction Project Management Knowledge Graph
CRCEE	Chinese Registered Construction Engineer Examination
GELU	Gaussian Error Linear Unit
GLM	General-purpose Large-language Model
MAMCQ	Multiple Answers Multiple-Choice Question
MTEB	Massive Text Embedding Benchmark
NLP	Natural Language Processing
RAG	Retrieval Augmented Generation
SAMCQ	Single Answer Multiple-Choice Question
SQL	Structured Query Language

CRediT authorship contribution statement

Shenghua Zhou: Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Keyan Liu:** Writing – original draft, Methodology. **Dezhi Li:** Writing – review & editing, Validation, Supervision, Conceptualization. **Chun Fu:** Validation, Supervision, Conceptualization. **Yan Ning:** Writing – review & editing, Formal analysis, Conceptualization. **Wenying Ji:** Writing – review & editing, Conceptualization. **Xuefan Liu:** Methodology, Data curation. **Bo Xiao:** Writing – review & editing, Conceptualization. **Ran Wei:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The team would like to thank National Natural Science Foundation of China (No. 72201057), Social Science Foundation of Jiangsu Province (No. 23GLC020), and China Scholarship Council (No. 202306090050) for financially supporting this work. All supplementary resources (i.e., demonstration videos, CPM-KG, CPM-QA testing set, detailed results, and codes) are freely available via HuggingFace (https://huggingface.co/datasets/AnonymousSite/QA_dataset_for_CPM) and GitHub (https://github.com/0AnonymousSite0/QA_for_CPM) repositories.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.aei.2025.103142>.

References

- [1] Ershadi, Goodarzi, Core capabilities for achieving sustainable construction project management, *Sustainable Prod. Consumption* 28 (2021) 1396–1410, <https://doi.org/10.1016/j.spc.2021.08.020>.
- [2] Laffaj, Rebai, Albalkhy, Hamdi, Mossman, Alves Da Costa, Complexity in construction projects: a literature review, *Buildings* 14 (3) (2024) 680, <https://doi.org/10.3390/buildings14030680>.
- [3] Expert Market Research, Global Construction Market Outlook, 2023, Retrieved from <https://www.expermarketresearch.com/reports/construction-market>.
- [4] Baviskar, Ahirrao, Poidar, Kotecha, Efficient automated processing of the unstructured documents using artificial intelligence: a systematic literature review and future directions, *IEEE Access* 9 (2021) 72894–72936, <https://doi.org/10.1109/ACCESS.2021.3072900>.
- [5] L. Paul, H. O'cass, Bresciani, Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR), *Int. J. Consum. Stud.* 45 (4) (2021) O1–O16, <https://doi.org/10.1111/ijcs.12695>.
- [6] T. Saka, S. Saka, A. Ajayi, Kazemi, GPT models in construction industry: Opportunities, limitations, and a use case validation, *Dev. Built Environ.* 17 (2023) 100300, <https://doi.org/10.1016/j.dbe.2023.100300>.
- [7] Jin, Zhao, Yu, Tao, Yin, Liu, Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning, *Data Min. Knowl. Disc.* 37 (1) (2023) 255–288, <https://doi.org/10.1007/s10618-022-00891-8>.
- [8] Zhu, Lei, Wang, Zheng, Chua, Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering, 2021, *arXiv preprint arXiv:2101.00774*. Doi: <https://doi.org/10.48550/arXiv.2101.00774>.
- [9] Fitkau, Hartmann, An ontology-based approach of automatic compliance checking for structural fire safety requirements, *Adv. Eng. Inf.* 59 (2024), <https://doi.org/10.1016/j.aei.2023.102314>.
- [10] Yin, Tang, Webster, Xu, Li, Ying, An ontology-aided, natural language-based approach for multi-constraint BIM model querying, *J. Build. Eng.* 76 (2023), <https://doi.org/10.1016/j.jobe.2023.107066>.
- [11] M. Hassan, Marengo, Nutt, A BERT-Based Model for Question Answering on Construction Incident Reports, Springer, Cham, 2022, pp. 215–223.
- [12] Zhang, El-Gohary, Natural language generation and deep learning for intelligent building codes, *Adv. Eng. Inf.* 52 (2022) 101557, <https://doi.org/10.1016/j.aei.2022.101557>.
- [13] Z. Ren, C. Li, L. Tian, Liu, Automatic quality compliance checking in concrete dam construction: Integrating rule syntax parsing and semantic distance, *Adv. Eng. Inf.* 60 (2024), <https://doi.org/10.1016/j.aei.2024.102409>.
- [14] Banerjee, Potts, Jhala, Jaselskis, Developing a construction domain-specific artificial intelligence language model for NCDOT's CLEAR program to promote organizational innovation and institutional knowledge, *J. Comput. Civ. Eng.* 37 (3) (2023), <https://doi.org/10.1061/JCCEES.CPENG-4868>.
- [15] S. Mohamed, Abourizk, Context-driven ontology-based risk identification for onshore wind farm projects: A domain-specific approach, *Adv. Eng. Inf.* 56 (2023) 101962, <https://doi.org/10.1016/j.aei.2023.101962>.
- [16] O. Saka, G. Akanbi, Chan, Bello, Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities, *Adv. Eng. Inf.* 55 (2023) 101869, <https://doi.org/10.1016/j.aei.2022.101869>.
- [17] Delong, Mir, Whyte, Ji, Fleuriot, Neurosymbolic AI for reasoning on graph structures: a survey, 2023, arXiv preprint arXiv: 2302.
- [18] Ko, Jeong, Lee, Natural Language Processing-Driven Model to Extract Contract Change Reasons and Altered Work Items for Advanced Retrieval of Change Orders, *J. Constr. Eng. Manag.* 147 (11) (2021) 04021147, [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002172](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002172).
- [19] Feng, Chen, A small samples training framework for deep Learning-based automatic information extraction: case study of construction accident news reports analysis, *Adv. Eng. Inf.* 47 (2021), <https://doi.org/10.1016/j.aei.2021.101256>.
- [20] Li, Mo, Yang, Li, Jiang, Wang, Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model, *Adv. Eng. Inf.* 50 (2021) 101416, <https://doi.org/10.1016/j.aei.2021.101416>.
- [21] Hu, Chapman, Wen, Hall, What can knowledge bring to machine learning?—a survey of low-shot learning for structured data, *ACM Transactions on Intelligent Systems and Technology (TIST)*. 13 (3) (2022) 1–45, <https://doi.org/10.1145/3510030>.
- [22] Wilkho, Chang, Gharaibeh, FF-BERT: a BERT-based ensemble for automated classification of web-based text on flash flood events, *Adv. Eng. Inf.* 59 (2024) 102293, <https://doi.org/10.1016/j.aei.2023.102293>.
- [23] T. Trappey, Wu, Wang, Intelligent compilation of patent summaries using machine learning and natural language processing techniques, *Adv. Eng. Inf.* 43 (2020) 101027, <https://doi.org/10.1016/j.aei.2019.101027>.
- [24] Hadi, Qureshi, Shah, Irfan, Zafar, Shaikh, Akhtar, Wu, Mirjalili, A survey on large language models: Applications, challenges, limitations, and practical usage, *Authorea Preprints*, 2023.
- [25] Zhu, Wang, Chen, Qiao, Ou, Yao, Deng, Chen, Zhang, LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities, *arXiv preprint arXiv:2305.13168*, 2023, Doi: [10.48550/arXiv.2305.13168](https://doi.org/10.48550/arXiv.2305.13168).
- [26] Yang, Jin, Tang, Han, Feng, Jiang, Zhong, Yin, Hu, Harnessing the power of LLMs in practice: a survey on chatgpt and beyond, *ACM Trans. Knowl. Discov. Data* (2023), <https://doi.org/10.1145/3649506>.
- [27] Xu, Qi, Wang, Wang, Zhang, Xu, Knowledge Conflicts for LLMs: A Survey, *arXiv preprint arXiv:2403.08319*, 2024, Doi: [10.48550/arXiv.2403.08319](https://doi.org/10.48550/arXiv.2403.08319).
- [28] Zafar, Sahoo, Bhardawaj, Das, Ekbal, KI-MAG: a knowledge-infused abstractive question answering system in medical domain, *Neurocomputing* 571 (2024), <https://doi.org/10.1016/j.neucom.2023.127141>.
- [29] I. Wang, Anumba, Query answering system for building information modeling using BERT NN Algorithm and NLG, *Computing in Civil Engineering* 2021 (2021) 425–432.
- [30] Ray, Chakrabarti, A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis, *Appl. Comput. Inf.* 18 (1/2) (2022) 163–178, <https://doi.org/10.1016/j.aci.2019.02.002>.
- [31] Wu, Wang, Wang, He, Zhu, Question Answering (QA) Basics. *Visual Question Answering: From Theory to Application*. Springer, 2022b, pp. 27–31. Doi: [10.1007/978-981-19-0964-1_3](https://doi.org/10.1007/978-981-19-0964-1_3).
- [32] Eligüzel, Çetinkaya, Dereli, Application of named entity recognition on tweets during earthquake disaster: a deep learning-based approach, *Soft. Comput.* 26 (1) (2022) 395–421, <https://doi.org/10.1007/s00500-021-06370-4>.
- [33] Deng, Chen, Huang, Chen, Cheng, Research on the construction of event logic knowledge graph of supply chain management, *Adv. Eng. Inf.* 56 (2023) 101921, <https://doi.org/10.1016/j.aei.2023.101921>.
- [34] Wang, Liu, Jia, Guan, Liu, Duan, Deep fuzzy rule-based classification system with improved wang-mendel method, *IEEE Trans. Fuzzy Syst.* 30 (8) (2021) 2957–2970, <https://doi.org/10.1109/TFUZZ.2021.3098339>.
- [35] Xin, Li, Li, Li, Zhou, Evaluations and comparisons of rule-based and machine-learning-based methods to retrieve satellite-based vegetation phenology using MODIS and USA National Phenology Network data, *Int. J. Appl. Earth Obs. Geoinf.* 93 (2020) 102189, <https://doi.org/10.1016/j.jag.2020.102189>.
- [36] Gao, Xiong, Gao, Jia, Pan, Bi, Dai, Sun, Wang, Retrieval-augmented generation for large language models: survey, 2023, arXiv preprint arXiv:2312.10997.
- [37] Wang, Lo, Text mining approaches for dealing with the rapidly expanding literature on COVID-19, *Brief. Bioinform.* 22 (2) (2021) 781–799, <https://doi.org/10.1093/bib/bbaa296>.
- [38] Xu, Cai, Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure, *Adv. Eng. Inf.* 48 (2021) 101288, <https://doi.org/10.1016/j.aei.2021.101288>.
- [39] Wu, Lin, Leng, Li, Hu, Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web, *Autom. Constr.* 135 (2022) 104108, <https://doi.org/10.1016/j.autcon.2021.104108>.
- [40] Su, Ai, Wu, Ma, Li, Liu, Caseformer: Pre-training for legal case retrieval, 2023, arXiv preprint arXiv:2311.00333. Doi: [10.48550/arXiv.2311.00333](https://doi.org/10.48550/arXiv.2311.00333).
- [41] Y. Lucas, Pomeroy, Yang, Reasoning with large language models for medical question answering, *J. Am. Med. Inform. Assoc.* 31 (9) (2024) 1964–1975, <https://doi.org/10.1093/jamia/ocae131>.

- [42] Wang, Yang, Qiu, Liang, He, Gu, Xiao, Wang, Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases, 2023, arXiv preprint arXiv:2308.11761. Doi: [10.48550/arXiv.2308.11761](https://doi.org/10.48550/arXiv.2308.11761).
- [43] Monir, Lau, Yang, Zhao, VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search, 2024, arXiv preprint arXiv: 2409.17383. Doi: [10.48550/arXiv.2409.17383](https://doi.org/10.48550/arXiv.2409.17383).
- [44] Chen, Xiao, Zhang, Luo, Lian, Liu, M3-embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024, pp. 2318-2335.
- [45] L. Sarhan, Portmann, Towards a Standard Feature Set for Network Intrusion Detection System Datasets, *Mobile Networks Appl.* 27 (1) (2022) 357–370, <https://doi.org/10.1007/s11036-021-01843-0>.
- [46] Longpre, Lu, Daiber, MKQA: a linguistically diverse benchmark for multilingual open domain question answering, *Transactions of the Association for Computational Linguistics*. 9 (2021) 1389–1406, https://doi.org/10.1162/tacl_a_00433.
- [47] Cremer, Sheehan, Fortmann, Kia, Mullins, Murphy, Materne, Cyber risk and cybersecurity: a systematic review of data availability, *The Geneva Papers on Risk and Insurance-Issues and Practice* 47 (3) (2022) 698–736, <https://doi.org/10.1057/s41288-022-00266-6>.
- [48] M. Raiaan, F.H. Fatema, M. Sakib, A. Ahmad, Azam, A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges, *IEEE Access* (2024), <https://doi.org/10.1109/ACCESS.2024.3365742>.
- [49] Yuan, Coenen, Reif, Ippolito, Wordcraft: Story Writing With Large Language Models [M], in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 841-852.
- [50] K. Madani, S. Greene, H. Mohr, X. Olmos, S. Sun, Fraser, Naik, Large language models generate functional protein sequences across diverse families, *Nat. Biotechnol.* 41 (8) (2023) 1099–1106, <https://doi.org/10.1038/s41587-022-01618-2>.
- [51] Frieder, Pinchetti, Griffiths, Salvatori, Lukasiewicz, Petersen, Berner, Mathematical capabilities of chatgpt, *Adv. Neural Inf. Proces. Syst.* 36 (2024).
- [52] Singhal, Tu, Gottweis, Sayres, Wulczyn, Hou, Clark, Pfohl, Cole-Lewis, Neal, Towards expert-level medical question answering with large language models, 2023b, arXiv preprint arXiv:2305.09617. Doi: [10.48550/arXiv.2305.09617](https://doi.org/10.48550/arXiv.2305.09617).
- [53] Gutiérrez-Fandiño, Armengol-Estabé, Pàmies, Llop-Palao, Silveira-Ocampo, Carrino, González-Agirre, Armentano-Oller, Rodríguez-Penagos, Villegas, María: Spanish language models, 2021, arXiv preprint arXiv:2107.07253. Doi: [10.26342/2022-68-3](https://doi.org/10.26342/2022-68-3).
- [54] Zhu, Qi, *Construction project management*, Southeast University Press, Nanjing, 2019.
- [55] Ding, Wang, Liu, Mao, Kong, *Construction project management 1Z200000, 2023 edition*, China Construction Industry Press, Beijing, 2023.
- [56] Yang, *Construction project management*, Wuhan University Press, Wuhan, 2023.
- [57] F. Rizaldy, Hartono, Systematic Literature Review: Current Products, Topic, and Implementation of Graph Database, *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*. 7 (1) (2021), <https://doi.org/10.26555/jiteki.v7i1.19007>.
- [58] Friis, Getting Started with langchain, 2023, Retrieved from <https://github.com/langchain-ai/langchain>.
- [59] Guan, Liu, Lin, Lu, He, Han, Sun, Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-Based Retrofitting, *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (16) (2024) 18126–18134, <https://doi.org/10.1609/aaai.v38i16.29770>.
- [60] Chang, Yu, Chang, Yang, Kumar, Pre-training tasks for embedding-based large-scale retrieval, 2020, arXiv preprint arXiv:2002.03932. Doi: [10.48550/arXiv.2002.03932](https://doi.org/10.48550/arXiv.2002.03932).
- [61] Massive Text Embedding Benchmark. 2024. Massive Text Embedding Benchmark Leaderboard. Retrieved from <https://huggingface.co/spaces/mteb/leaderboard>.
- [62] Beltagy, Peters, Cohan, Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020, Doi: [10.48550/arXiv.2004.05150](https://doi.org/10.48550/arXiv.2004.05150).
- [63] Yang, Zhang, Li, Bendersky, Najork, Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching, 2020, 1725-1734. Doi: [10.1145/3340531.3411908](https://doi.org/10.1145/3340531.3411908).
- [64] Koh, Ju, Liu, Pan, An empirical survey on long document summarization: Datasets, models, and metrics, *ACM Comput. Surv.* 55 (8) (2022) 1–35, <https://doi.org/10.1145/3545176>.
- [65] Menghani, Efficient deep learning: a survey on making deep learning models smaller, faster, and better, *ACM Comput. Surv.* 55 (12) (2023) 1–37, <https://doi.org/10.1145/3578938>.
- [66] Dmetasoul, Introduction of Dmeta-embedding-zh, 2024, Retrieved from [htt ps://huggingface.co/DMetaSoul/Dmeta-embedding-zh](https://huggingface.co/DMetaSoul/Dmeta-embedding-zh).
- [67] Nan, Sheng, Cao, Hu, Wang, Li, Let silence speak: Enhancing fake news detection with generated comments from large language models, 2024, 1732-1742. Doi: [10.1145/3627673.3679519](https://doi.org/10.1145/3627673.3679519).
- [68] Cao, Gio, Xu, Robust damage localization in plate-type structures by using an enhanced robust principal component analysis and data fusion technique, *Mech. Syst. Sig. Process.* 162 (2022) 108091, <https://doi.org/10.1016/j.ymssp.2021.108091>.
- [69] Zhang, Li, Wang, Wu, A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM, *Measurement* 173 (2021) 108644, <https://doi.org/10.1016/j.measurement.2020.108644>.
- [70] Jung, Jung, Kim, Won, Mun, Classifying the risk of cognitive impairment using sequential gait characteristics and long short-term memory networks, *IEEE J. Biomed. Health Inform.* 25 (10) (2021) 4029–4040, <https://doi.org/10.1109/JBHI.2021.3073372>.
- [71] Fujian Personnel Testing Authority, Public announcement of qualified persons of Fujian examination area for the 2023 qualification examination for first-level registered constructors, 2024, Retrieved from <https://www.fjpta.com/portal/indexbulletin/bGWypfh.htm>.
- [72] Bouabdil, Kim, Ermis, Fadaee, Hooker, Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation, 2023, arXiv preprint arXiv: 2310.14424. Doi: [10.48550/arXiv.2310.14424](https://doi.org/10.48550/arXiv.2310.14424).
- [73] Jiang, Ren, Lin, Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023, arXiv preprint arXiv:2306.02561. Doi: [10.48550/arXiv.2306.02561](https://doi.org/10.48550/arXiv.2306.02561).
- [74] Zhang, Zhang, Yuan, Liu, Shi, Gui, Zhang, Huang, Llmeval: A preliminary study on how to evaluate large language models, 2024, pp. 19615-19622. Doi: [10.1609/aaai.v38i17.29934](https://doi.org/10.1609/aaai.v38i17.29934).
- [75] Zhong, Wu, Xiang, Guo, Automatic information extraction from construction quality inspection regulations: a knowledge pattern-based ontological method, *J. Constr. Eng. Manag.* 148 (3) (2022) 04021207, [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002240](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002240).
- [76] Peng, Liu, Yang, Yuan, Li, Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data, 2023, arXiv preprint arXiv:2308.03107. Doi: [10.48550/arXiv.2308.03107](https://doi.org/10.48550/arXiv.2308.03107).
- [77] Jia, Jia, Kong, Lin, Qi, Privacy-aware retrieval of electronic medical records by fuzzy keyword search, *Hum Centric Comput Inf Sci.* 12 (2022) 1–15, <https://doi.org/10.22967/HCIS.2022.12.041>.
- [78] Cheng, Yang, Xiang, Liu, Financial time series forecasting with multi-modality graph neural network, *Pattern Recognit.* 121 (2022) 108218, <https://doi.org/10.1016/j.patcog.2021.108218>.
- [79] Zeng, Tu, Liu, Fu, Su, Toward better drug discovery with knowledge graph, *Curr. Opin. Struct. Biol.* 72 (2022) 114–126, <https://doi.org/10.1016/j.sbi.2021.09.003>.
- [80] Wu, Irsøy, Lu, Dabrowski, Dredze, Gehrmann, Kambadur, Rosenberg, Mann, Bloomberggpt: A large language model for finance, 2023, arXiv preprint arXiv: 2303.17564. Doi: [10.48550/arXiv.2303.17564](https://doi.org/10.48550/arXiv.2303.17564).
- [81] A. Singhal, M. Tu, C. Wei, T. Scales, Cole-Lewis, Pfohl, Large language models encode clinical knowledge, *Nature* 620 (7972) (2023) 172–180, <https://doi.org/10.48550/arXiv.2212.13138>.
- [82] Boiko, Macknight, Gomes, Emergent autonomous scientific research capabilities of large language models, 2023, arXiv preprint arXiv:2304.05332. Doi: [10.48550/arXiv.2304.05332](https://doi.org/10.48550/arXiv.2304.05332).
- [83] Fergus, Botha, Ostovar, Evaluating academic answers generated using ChatGPT, *J. Chem. Educ.* 100 (4) (2023) 1672–1675, <https://doi.org/10.1021/acs.jchemed.3c00087>.