

A cultural-element-augmented framework for Chinese culture-loaded words related question answering

Fangfang Xu^{1,2} , Wei Zhang^{1,2}, Wei Tu³, Haijiang Liu^{1,2}, Qiaoling Xiao^{4*} and Jinguang Gu^{1,2*} 

¹ School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China

² Hubei Provincial Key Laboratory of Intelligent Information Processing and Real-time Industrial Systems, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China

³ Network and Information Center, Wuhan University of Science and Technology, Wuhan, Hubei 430081, China

⁴ School of International Education, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China

* Correspondence: christina.xiao@wust.edu.cn (Xiao Q); simon@ontoweb.wust.edu.cn (Gu J)

Abstract

Understanding and answering questions about Chinese culture-loaded words that embody their connotations can help large language models better support cultural education for adolescents, cultural content creation, and intercultural communication. However, due to limited coverage of cultural-domain knowledge, large language models are prone to factual errors and content hallucinations in culturally situated question answering. Although using a knowledge graph (KG) can improve accuracy and reliability, in the strongly context-dependent setting of culture-loaded words, existing general-purpose KG-LLM approaches still face three key challenges: (1) entity linking lacks constraints from cultural elements, expanding the candidate space and introducing irrelevant paths; (2) path retrieval and pruning are not guided by cultural semantic information and rely excessively on the LLM's judgments, causing erroneous expansion/pruning and noise; and (3) the ranking of different cultural evidence paths is inaccurate, hindering consistent selection of high-confidence evidence and undermining answer quality and explanation completeness. To address these issues, this paper proposes a cultural-element-augmented KG-LLM question-answering framework for Chinese culture-loaded words (CEAF-CLWQA). CEAF-CLWQA extracts and organizes cultural-element constraints using prompt templates for entity linking, retrieves candidate paths under cultural-element and structural constraints with cultural semantic guidance for retrieval and pruning, and ranks cultural evidence to generate traceable and explainable natural-language answers based on the Top-K evidence paths. We constructed a Chinese culture-loaded words question-answering dataset containing 100,216 question-answer pairs; experiments on three LLM backbones show that CEAF-CLWQA achieves the best Hits@1 across settings and improves F1 over strong KG-LLM baselines (e.g., ToG and PoG) on the two stronger backbones, reaching 87.1% Hits@1 and 68.3% F1 with DeepSeek-V3.

Keywords: Culture-loaded words, Knowledge graph question answering, Large language models

Citation: Xu F, Zhang W, Tu W, Liu H, Xiao Q, et al. 2026. A cultural-element-augmented framework for Chinese culture-loaded words related question answering. *The Knowledge Engineering Review* 41: e015 <https://doi.org/10.48130/ker-0026-0011>

1 Introduction

Culture-loaded words refer to words, phrases, and idioms that denote things unique to a particular culture; they reflect the distinctive ways of life that a given people have gradually accumulated over a long historical process, in ways that differ from those of other peoples^[1]. Understanding and answering questions about culture-loaded words that carry cultural connotations is of significant value for adolescents' cultural education, cultural content creation, and cultural dissemination and communication^[2,3]. Unlike open-domain question answering, culture-loaded words question answering typically concerns the cultural connotations of a term and its origins, requiring the model to capture cultural knowledge, including allusions, sources, quotation contexts, and interpretations.

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across a variety of natural language processing tasks^[4–6]. However, when queries concern knowledge-intensive and strongly context-dependent domains, the limitations of LLMs become increasingly evident: with insufficient coverage of domain knowledge, models are prone to factual errors and content hallucinations^[7–10]. This issue substantially constrains the application of LLMs in cultural question-answering scenarios that demand high reliability and factual accuracy.

To mitigate hallucinations and factual errors, incorporating a knowledge graph (KG) as an external, verifiable source of facts to augment LLMs' reasoning has become a critical approach^[11–14]. A KG organizes entities, relations, and type information in a structured form, providing a traceable evidential basis for multi-hop reasoning. In the domain of Chinese culture-loaded words, content-rich knowledge graphs have been successfully constructed, laying a data foundation for a systematic understanding of Chinese culture. However, how to efficiently integrate such structured domain knowledge with language understanding and reasoning capabilities of LLMs remains challenging, and existing general-purpose KG-LLM methods still face three key challenges in the strongly context-dependent setting of culture-loaded words:

1. Entity linking lacks constraints from cultural elements: Culture-loaded words question answering is strongly context-dependent, and questions often implicitly contain multiple cultural elements that constitute knowledge constraints. Without explicit constraints on these cultural elements, it is difficult in subsequent retrieval to precisely determine entity boundaries and the candidate scope consistent with the question context, causing the search space to expand substantially and allowing a large number of irrelevant paths to enter the candidate set, thereby increasing retrieval and filtering difficulty and degrading the final reasoning quality.

2. Path retrieval and pruning lack guidance from cultural semantic information: In existing general-purpose KG–LLM methods, stages such as path expansion, pruning, and candidate filtering are often determined directly by the LLM, with little guidance from cultural semantic information. When the LLM has insufficient coverage of the target cultural domain, the retrieval process is likely to introduce irrelevant paths or mistakenly prune crucial evidence paths for cultural interpretation, thereby accumulating noise and reducing the reliability of reasoning.

3. Inaccurate ranking of different cultural evidence paths: The retrieved knowledge may contain multiple candidate paths; these cultural evidence paths differ in their degree of alignment with the question context, their evidential strength, and their information coverage. Without a ranking mechanism tailored to cultural evidence, it is difficult to consistently identify higher-confidence evidence paths, thereby affecting the accuracy and overall quality of the final answers.

Building on the above challenges, this paper focuses on the following research questions:

1. How can cultural elements in culture-loaded words questions be leveraged and transformed into a controllable constraint representation to constrain the subsequent retrieval process?
2. How can cultural semantic information be used to guide path retrieval and pruning, to suppress irrelevant path noise while improving candidate coverage and reducing reliance on the LLM's subjective judgments?
3. How can different candidate paths be more accurately ranked and filtered with respect to cultural evidence, and how can high-quality answers be generated based on the Top-K evidence set?

To address the above questions, this paper proposes a cultural-element-augmented KG–LLM question-answering framework for Chinese culture-loaded words (CEAF-CLWQA). The core idea is to explicitly incorporate constraints from cultural elements, cultural semantic information, and a cultural evidence selection mechanism into key decision-making steps, and to apply progressive constraints via 'cultural-element-based entity linking—path retrieval and pruning guided by cultural semantic information—cultural-evidence-based path ranking', thereby mitigating retrieval drift caused by unstable entity alignment, reducing reliance on the LLM's subjective judgments in critical stages, and supporting accurate answer generation with a Top-K set of cultural evidence. Specifically, the framework consists of four phases: first, it uses an LLM to extract and organize cultural-element constraints from the user's question, and performs entity linking to the knowledge graph; second, it retrieves candidate paths under constraints from cultural elements and structural constraints, and incorporates cultural semantic information to guide retrieval and pruning, improving candidate coverage while suppressing irrelevant noise; then, it ranks candidate paths based on cultural elements and stably selects the best-matching Top-K evidence paths from the candidate set; finally, it generates a more information-complete natural-language answer grounded in the Top-K cultural evidence paths. Experimental results show that, across three LLM backbones, our framework achieves the best performance in terms of Hits@1, and it also outperforms a range of state-of-the-art baselines—including ToG and PoG—on F1 when using the two stronger backbones.

The main contributions of this paper are summarized as follows:

1. We propose a cultural-element-augmented KG–LLM framework for question answering on Chinese culture-loaded words. Through a progressive constraint mechanism—'cultural-element-based entity linking—path retrieval and pruning guided by cultural semantic information—cultural-evidence-based path ranking'—the framework explicitly injects cultural elements and cultural semantic cues into key

decision-making steps, thereby improving the controllability and reliability of the question-answering process.

2. We design a cultural-semantic-information-guided strategy for path retrieval and pruning, and propose a cultural-evidence-based path ranking method. These components enable stable selection while maintaining candidate coverage, effectively suppressing the accumulation of irrelevant-path noise and reducing reliance on the LLM's subjective judgments in critical stages.

3. We construct a Chinese culture-loaded words question-answering dataset containing 100,216 question–answer pairs. Through systematic comparisons with LLM-only methods (Zero-Shot and CoT) and representative KG–LLM methods (e.g., ToG and PoG), experiments validate the advantages of our framework in terms of question answering accuracy, reliability, traceability, and explainability, providing a reusable technical solution for knowledge-enhanced question answering in the cultural domain.

2 Related work

2.1 Cultural-domain question answering

The concept of culture encompasses the distinct ways in which different groups think, feel, and behave, thereby distinguishing them from one another^[15]. In recent years, researchers have proposed numerous culture-related QA tasks to assess large language models' cultural knowledge, thereby evaluating their performance in scenarios with high cultural knowledge demands^[16–18]. Representative efforts include CulturalBench, which covers multiple regions and topics^[19]; CaLMQA, a multilingual benchmark for culture-specific long-form question answering^[20]; BertaQA, which contrasts local cultural knowledge with globally common knowledge^[21]; and AraDiCE-Culture, which focuses on fine-grained regional cultural differences^[22]. In the Chinese context, related evaluations and benchmarks also include CHARM^[23], HalluQA^[24], and CIF-Bench^[25], among others.

The above studies indicate—from perspectives such as cross-regional/cross-lingual transfer, coverage of culture-specific knowledge, and reasoning consistency in cultural contexts—that compared with culture-agnostic or general-topic question answering, cultural-domain question answering is more likely to expose models' knowledge gaps and contextual misinterpretations, thereby causing more frequent factual errors and hallucinations.

2.2 Knowledge organization in the cultural domain

To mitigate factual errors and hallucinations that large language models often produce in cultural question answering, prior work has focused on the structured organization of cultural-domain knowledge, particularly on the construction and application of cultural-domain knowledge graphs. Standardized knowledge organization frameworks such as CIDOC-CRM provide a foundation for semantic interoperability and federated querying of cross-institutional cultural data^[26], and have inspired a variety of cultural heritage knowledge graph initiatives, such as ArCo^[27] and WarSampo^[28]. In the Chinese context, researchers have also explored knowledge graph construction and knowledge services for cultural resources such as intangible cultural heritage, for example, Nanjing Yunjin^[29] and ancient costume culture^[30].

However, most of the above knowledge graphs focus on specific cultural scenarios and applications, with narrow cultural content coverage. As a result, they fail to capture the connections among different types of cultural knowledge. They cannot provide a more comprehensive and stable knowledge source, making it difficult to further reduce

hallucinations and factual errors in question answering. In contrast, the culture-loaded words studied in this paper define lexical-level cultural knowledge, covering cultural content across multiple domains, including history, humanistic traditions, social customs, and cultural heritage. Therefore, organizing knowledge about culture-loaded words can provide large language models with retrievable factual grounding and traceable evidence chains, thereby improving answer accuracy and reducing the risk of factual errors and hallucinations in cultural-context QA. Zhang et al.^[31] constructed a Chinese culture-loaded words knowledge graph and used an extensible ontology to structurally organize the core knowledge dimensions of Chinese culture-loaded words and their associated relations.

However, existing studies have primarily focused on knowledge organization, and there has been limited research on more complex culture-loaded words question answering. Such question answering is required not only to provide the meaning of a cultural concept but also to explain its cultural connotations, contextual sources, and other related knowledge. To fill this gap, this paper focuses on question answering for Chinese culture-loaded words. It investigates how a Chinese culture-loaded word KG can be leveraged to enhance the performance of large language models on this task.

2.3 KG-enhanced large language model question answering frameworks

Generic KG-enhanced large language model approaches typically involve the following steps: (1) question parsing/entity linking; (2) subgraph (or relevant KG knowledge) retrieval; and (3) injecting structured information such as triples/paths into prompts to generate answers^[32].

In the question parsing/entity linking stage, KAPING^[33] identifies question entities and retrieves relevant triples, which are then concatenated into the prompt to enable zero-shot KGQA. In the subgraph retrieval stage, ToG^[34] performs a beam search over the graph to obtain multi-hop evidence, PoG^[35] improves path relevance through structured path exploration and multi-strategy pruning, and StructGPT^[36] enhances the use of structured evidence via an iterative 'reading structured evidence—reasoning (IRR)' process. RoG^[37] first generates a KG-constrained relation-path plan, then retrieves evidence accordingly, and produces answers with explanations. In the information injection and generation stage, MindMap^[38] leverages knowledge graph prompting to enhance structured reasoning and explainability.

Although the above methods have achieved substantial progress on general-purpose knowledge graphs (e.g., Wikidata/Freebase) and standard KGQA benchmarks, directly transferring them to the culture-loaded words question answering setting still suffers from issues such as entity linking without constraints from cultural elements, path retrieval and pruning without guidance from cultural semantic information, and inaccurate ranking of different cultural evidence paths.

To this end, we propose a cultural-element-augmented KG-LLM framework for question answering on Chinese culture-loaded words. The framework uses cultural elements extracted from questions to guide KG retrieval and path pruning. It ranks the resulting cultural evidence paths according to cultural knowledge at different levels, thereby improving response accuracy and reducing hallucinations.

3 Overview of the Chinese culture-loaded words KG

Prior work models knowledge related to Chinese culture-loaded words primarily around culture-loaded words, context, and cultural

content categories: nodes for Chinese culture-loaded words record attributes such as orthography and pinyin, and cover characters, words, and sentences at different levels of expression granularity; context is used to characterize contextual evidence and is further subdivided into allusions, quotations, sources, and explanatory interpretations; cultural content categories adopt a two-level taxonomy of 'five upper-level categories + eight sub-level categories', facilitating aggregated retrieval along cultural dimensions^[31]; definitions of the five upper-level and eight sub-level categories are provided in [Supplementary Tables S1 and S2](#) in the Supplementary Information, and entity type definitions are given in [Supplementary Table S3](#). Built upon the above core concepts, the knowledge graph further operationalizes source information into provenance structures such as chapters, cultural sources, persons, and historical periods, thereby forming an evidence chain that enables tracing answers back to the original materials (the associations among entity types are illustrated in [Fig. 1](#)). The knowledge graph comprises 222,982 entities and 403,352 factual triples and provides rich coverage of culture-loaded words and their contextual relations. In this paper, cultural elements refer to a finite set corresponding to key entity types in the culture-loaded words knowledge graph ontology and their subtypes, which are used to represent the contextual and evidential dimensions required by a question and to constrain entity linking, path retrieval/pruning uniformly, and evidence ranking in subsequent stages. For example, for Records of the Grand Historian (Shiji), the corresponding entity types in the knowledge graph include the upper-level type 'cultural source' and its subtype 'book'; therefore, in our set of cultural elements, it can be categorized as 'cultural source' and 'book'.

4 Methodology

As illustrated in [Fig. 2](#), the overall architecture of the proposed framework comprises four core stages: **(1) Cultural-element-based entity linking**: this stage takes a natural-language question as input and extracts and organizes the cultural elements in the question via a cultural-element-guided prompt template; the resulting information is used to characterize the conditional entities, cultural-element constraints, and target elements, thereby forming a structured constraint representation that reduces semantic mismatch and narrows the candidate space. **(2) Path retrieval and pruning guided by cultural semantic information**: based on the structured constraint representation obtained in the previous stage, we perform constraint-based search over the knowledge graph to enumerate candidate reasoning paths, and then guide path expansion and pruning using the semantic information of cultural entities to obtain a set of potentially relevant knowledge paths, thereby suppressing irrelevant expansions that are inconsistent with the question context. **(3) Cultural-evidence-based path ranking**: the candidate-path set, while enhanced in coverage, may still contain noisy paths that are structurally feasible but insufficiently semantically relevant. To this end, this stage adopts a cultural-evidence-oriented ranking model to score and filter candidate paths, computing similarity scores between path nodes and the question via a learnable scoring model and jointly assessing how well each path, as cultural evidence, matches the question context. **(4) Answer generation**: we extract the Top-K cultural evidence paths that best match the question and normalize them into injectable evidence snippets (e.g., provenance, quotation context, and interpretations), which are then combined with the original question to construct a prompt for the LLM to generate a traceable, explainable, and more information-complete natural-language answer.

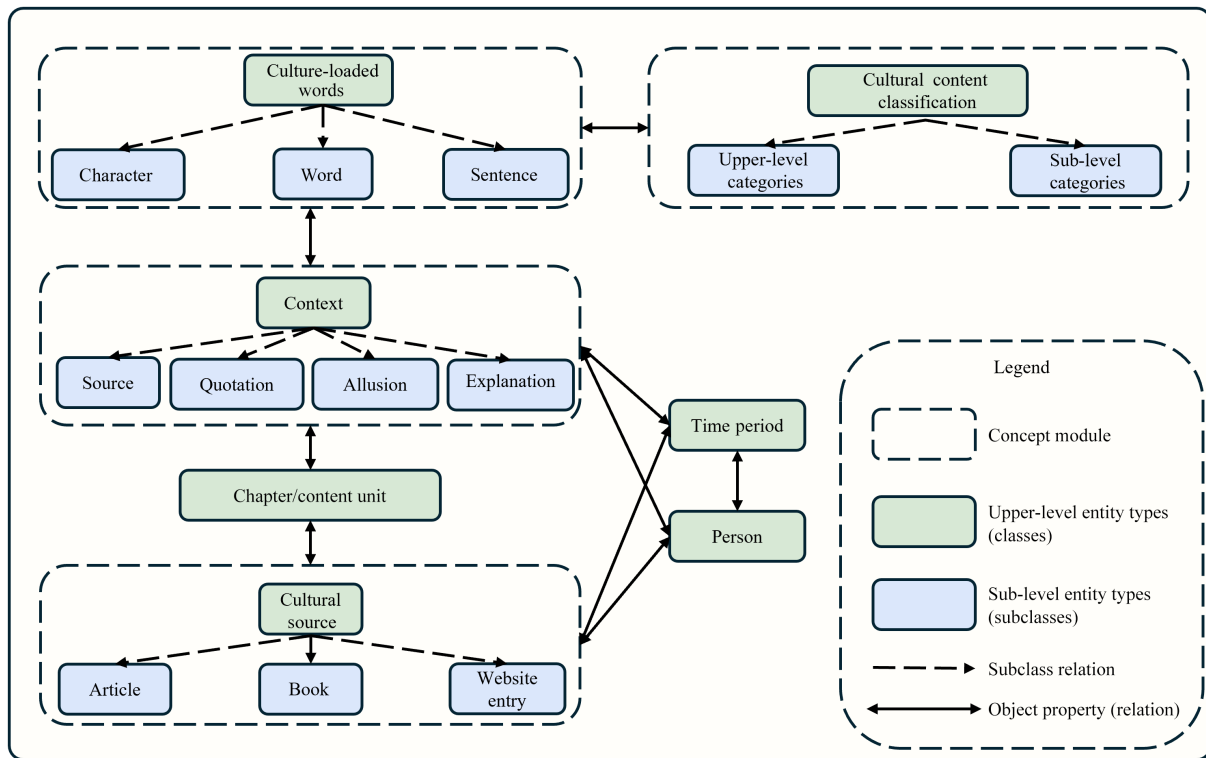


Fig. 1 Schematic diagram of the ontology structure of the Chinese culture-loaded words knowledge graph. The figure illustrates the hierarchical and object property relations among culture-loaded words, context, cultural content classification, and source chains.

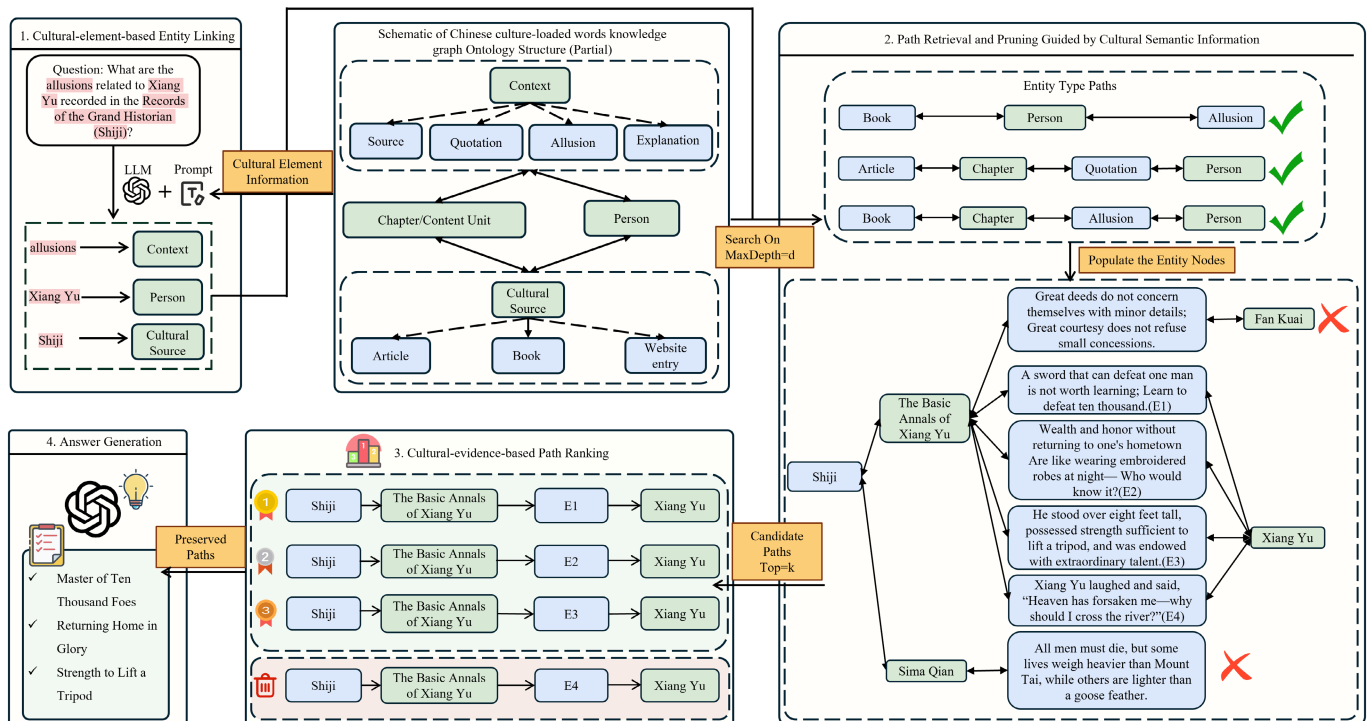


Fig. 2 Overall architecture of the proposed CEAF-CLWQA framework: an LLM first identifies cultural elements and links entities under constraints, then retrieves and prunes candidate knowledge-graph paths with cultural semantic guidance (within a maximum depth, d), ranks the resulting cultural evidence paths, and finally generates a traceable answer grounded in the selected evidence. An example question about the allusions related to Xiang Yu recorded in Shiji is shown to illustrate the end-to-end workflow.

4.1 Cultural-element-based entity linking

To transform natural-language questions into executable retrieval constraints and alleviate the ambiguity and missing constraints

commonly observed in culture-loaded words question answering, we construct a cultural-element-guided prompt template, as shown in Fig. 3, to extract and organize cultural elements from the question and



Fig. 3 A cultural-element-guided prompt template for entity linking.

generate a structured constraint representation, which includes the anchor entity, the cultural elements associated with the anchor entity, and the cultural elements associated with the target entity:

- **Anchor entity:** the key entity (one or more) explicitly mentioned in the question that can be used to locate the starting point of retrieval.
- **Cultural elements associated with the anchor entity:** the cultural elements that match the anchor entity, which constrain the semantic boundary and structural scope for entity alignment and subsequent retrieval.
- **Cultural elements associated with the target entity:** the cultural elements corresponding to the information sought by the question, which specify the element scope that subsequent retrieval needs to cover.

Using the above prompt template, we provide the model with the set of cultural elements defined in the knowledge graph along with their explanatory descriptions, the question text, and the constraint conditions, thereby guiding it to identify and extract the anchor entity from the question and to match the corresponding cultural elements for both the anchor entity and the target entity; the model ultimately outputs a structured constraint representation containing these three types of

information. Importantly, this stage is not an open-ended cultural inference based solely on the LLM's parametric knowledge. Instead, the extraction is explicitly constrained by three factors: (i) anchor entities are restricted to entities explicitly mentioned in the question; (ii) both anchor-element labels and the target-element label must be selected from the ontology-derived closed set T provided in the prompt; and (iii) explanatory descriptions of the candidate cultural elements are supplied to support label disambiguation. In addition, Prompt 1 requires exact label copying rather than paraphrasing or inventing new labels, and when multiple labels are ontologically plausible, a higher-level element is preferred to preserve downstream retrieval recall. These constraints narrow the space of admissible outputs at the very first stage and reduce the risk that an early cultural-element misidentification will propagate into large retrieval failures.

4.2 Path retrieval and pruning guided by cultural semantic information

Based on the structured constraint representation generated in the previous stage, this stage aims to retrieve, within a predefined

maximum search depth, as many cultural evidence paths relevant to the question as possible, thereby supporting subsequent path ranking and answer generation. Unlike unconstrained expansion directly on the instance graph, we decompose the retrieval process into two steps:

First, we retrieve entity-type paths that contain the cultural elements extracted in Stage 1, and then retrieve the corresponding entity paths on the knowledge graph conditioned on these entity-type paths. To obtain entity-type paths that satisfy the constraints, we adopt a multi-source breadth-first search. Because there may be multiple anchor entities, each of which may be associated with multiple cultural elements, we treat these cultural elements as starting nodes and expand them level by level. When an entity-type node is connected to an explored node, we include it in the expansion to form a longer entity-type path; expansion along a path stops once its length reaches the maximum search depth. After the search terminates, we discard entity-type paths that do not satisfy the constraints to reduce irrelevant expansions, thereby substantially shrinking the search space in the next step and filtering out many irrelevant entity paths.

Second, starting from the anchor entity, we populate the entity nodes that conform to the paths obtained in Step 1, based on the neighboring nodes and relations around the anchor entity. Because the search is multi-source, it produces many paths from different starting nodes; we therefore merge and deduplicate the retained paths to obtain the final set of complete candidate paths.

At this stage, pruning is mainly structural: we discard entity-type paths that do not satisfy the question-derived constraints before expanding them into concrete candidate paths. However, the retained candidate set may still contain structurally feasible but contextually weak paths. Therefore, in Stage 3, we further filter these candidates with the node scoring model to identify the most relevant evidence paths for answer generation.

4.3 Cultural-evidence-based path ranking

The candidate-path set obtained in the previous stage is often large and may include structurally feasible but weakly supported paths. This issue is particularly pronounced in cultural-domain question answering, where content with high surface-level similarity may correspond to different cultural elements. To prioritize the most relevant evidence, we rank candidate paths using a semantic-similarity model that incorporates cultural-element embeddings. Rather than relying solely on textual similarity, the model leverages cultural-element information to better distinguish genuinely supportive evidence from noisy alternatives, and we retain only the Top-K most relevant cultural evidence paths for answer generation.

4.3.1 Node scoring model and similarity label construction

The relevance between a candidate node and a question depends not only on textual semantics but also on the node's cultural-element-related evidential role in the knowledge graph. To train the node scoring model, we constructed an auxiliary dataset of question–node pairs from candidate evidence nodes in the knowledge graph. Rather than relying on unconstrained LLM judgments, we assign labels using a canonical five-level relevance rubric that reflects the question-conditioned evidential role of a node.

Specifically, for a question and a candidate node, the label of the pair is first assigned to one of five ordinal relevance bands: direct answer-bearing node (L5), close semantic match (L4), key provenance-support node (L3), auxiliary localization/support node (L2), and irrelevant node (L1). These bands are designed to encode ordinal evidential strength rather than a uniquely correct interval scale. For regression training, each band is deterministically mapped to a canonical representative soft-label value corresponding to the midpoint of

an equal-width relevance interval: 0.9 for L5, 0.7 for L4, 0.5 for L3, 0.3 for L2, and 0.1 for L1. Under this rubric, question–node pairs are assigned to one of the five bands and manually rechecked, and ambiguous cases are resolved by expert adjudication. The complete rubric and representative labeled examples are provided in [Supplementary Tables S4](#) and [S5](#), respectively, with further details on question-conditioned label assignment in [Supplementary Note 1](#).

Training objective: During training, we learn the model parameters in a supervised manner on the above question–node pairs using the mean squared error loss:

$$\mathcal{L}_{rank} = \frac{1}{N} \sum_{j=1}^N (s_j - y_j; \Theta)^2, \quad (1)$$

where N is the number of training samples, s_j denotes the predicted score of the j^{th} sample, y_j denotes the representative soft-label value associated with its assigned relevance band, and Θ represents the model parameters.

Predicted scoring: We first feed the question text and the node text into a pretrained text encoder and extract the [CLS] representation, denoted as h_{cls} . We then incorporate cultural-element information and obtain a node relevance score via fusion and regression:

(1) Cultural-element embedding: We learn a trainable embedding for each cultural element in the set T and maintain an embedding matrix $W_{type} \in \mathbb{R}^{|T| \times d_{type}}$, where d_{type} is the embedding dimension of cultural elements. For a node to be scored, we look up its embedding according to its cultural-element type:

$$h_{type} = W_{type} [type(\cdot)]. \quad (2)$$

(2) Similarity regression: we feed $h_{fused} = \text{Concat}(h_{cls}, h_{type})$ into a three-layer fully connected feed-forward network (MLP) and apply a sigmoid function to output the final similarity score $s \in [0, 1]$:

$$s = \sigma(MLP(h_{fused})). \quad (3)$$

The ranking module only reorders candidate nodes and paths retrieved from the knowledge graph. Therefore, its main failure mode is mis-ranking contextually weak but structurally feasible candidates, rather than introducing unsupported evidence outside the retrieved candidate space. Training dynamics and type-wise prediction behavior of the node scoring model are described in Section 5.4.

4.3.2 Path scoring and ranking

For the set of candidate paths, we score the cultural evidence nodes on each path and take the mean of the evidence-node scores within the path as the overall path score, thereby avoiding bias introduced by differences in path length. After computing the overall scores for all candidate paths, we rank them in descending order and select the k highest-scoring paths as the final set of cultural evidence paths used for answer generation.

4.4 Answer generation

After obtaining the top k highest-scoring cultural evidence paths through path ranking, this stage aims to transform structured evidence into user-facing natural-language answers and, under evidence constraints, provide necessary provenance and contextual support as much as possible, thereby improving the traceability and explainability of the answers while reducing the risks of factual errors and hallucinated content.

Specifically, we first linearize each evidence path and organize it into a set of numbered evidence snippets using a unified template. We then construct a prompt by combining the question with all evidence snippets, explicitly constraining the model to summarize and integrate information solely from the provided evidence and to annotate key

conclusions with the corresponding evidence indices or provenance identifiers, thereby producing a traceable natural-language answer.

5 Experiments

5.1 Dataset construction and statistics

Since there is currently no publicly available dataset tailored to culture-loaded words question answering, we construct a dataset for this domain by generating question-answer pairs in two ways:

(1) **Template-based generation:** We design a set of question-answer templates that cover different cultural elements and automatically generate a large number of single-hop and multi-hop questions and their answers by instantiating the templates with entities and relations in the knowledge graph; the question templates and subtype-specific instantiation examples are provided in [Supplementary Fig. S1](#) and [S2](#), respectively, with further details in [Supplementary Note 2](#).

(2) **LLM-based generation:** To complement questions with more diverse and natural formulations, we use the LLM in a knowledge-graph-grounded manner rather than as a free-form generator. For each selected culture-loaded word, we first execute ontology-guided Cypher queries over the knowledge graph to retrieve a local evidence subgraph. When the retrieved results involved quotation or provenance nodes, we further expand the subgraph to include explanation, chapter/content-unit, book, author, and time-period information when available. The retrieved subgraph is then serialized into JSON and provided to the LLM as structured evidence input for question generation, so that the generated questions and their associated answer/context information remained grounded in the retrieved graph evidence rather than in unconstrained parametric knowledge. The ontology-guided Cypher query is provided in [Supplementary Fig. S3](#).

After generation, we manually check 10% of the LLM-generated subset against the retrieved evidence, focusing on evidence faithfulness, cultural-connotation consistency, provenance/context accuracy, and question validity. We then assemble a dataset containing 100,216 question-answer pairs, including both single-hop and multi-hop questions, to enable a comprehensive evaluation across diverse cultural content and varying reasoning depths.

Data split. We split the 100,216 question-answer pairs into training/validation/test sets with an 80/10/10 ratio using stratified sampling

over (i) cultural categories, and (ii) reasoning-hop counts to preserve the original distribution. The validation set is used for early stopping and hyperparameter selection, and all results are reported on the held-out test set.

From the perspective of cultural content, we categorize and count all questions according to eight fine-grained cultural categories, yielding the distribution shown in [Fig. 4a](#). As can be seen, questions related to arts and literature are the most frequent in the dataset, totaling 26,500 (approximately 26.4%), followed by customs and practices (22,150, approximately 22.1%), and history and geography (18,400, approximately 18.4%). Questions in language and dialects and religion and beliefs account for 13,800 (approximately 13.8%) and 9,500 (approximately 9.5%), respectively. Although categories such as social institutions, emotions and values, and technical and specialized constitute smaller proportions (1.2%–5.2%), they still cover 1,166–5,200 questions, ensuring the visibility of long-tail cultural knowledge in the evaluation. Overall, each cultural dimension has a non-trivial scale, enabling systematic assessment of the model's question-answering capability across diverse cultural content types.

To characterize the reasoning difficulty of the question-answering task, we further categorize questions into four groups—1-hop, 2-hop, 3-hop, and 4+ hops—according to the length of the corresponding standard reasoning path, and report the proportion of each group as shown in [Fig. 4b](#). Overall, 1-hop questions account for approximately 45.0%, 2-hop questions for approximately 32.5%, 3-hop questions for approximately 15.3%, and questions requiring no fewer than four hops account for approximately 7.2%. Multi-hop (≥ 2 -hop) questions constitute more than half of the dataset, which enables a thorough evaluation of how our approach improves the quality of LLM-generated answers in complex cultural question-answering scenarios.

5.2 Evaluation metrics

To comprehensively evaluate question-answering performance, we adopt four standard metrics that are widely used in KGQA tasks:

- **Hits@1:** This metric measures whether the model's top-ranked answer is correct. For each question q , it is assigned a value of 1 if any answer in the gold answer set appears at the first position of the model's returned results; otherwise, it is assigned 0.

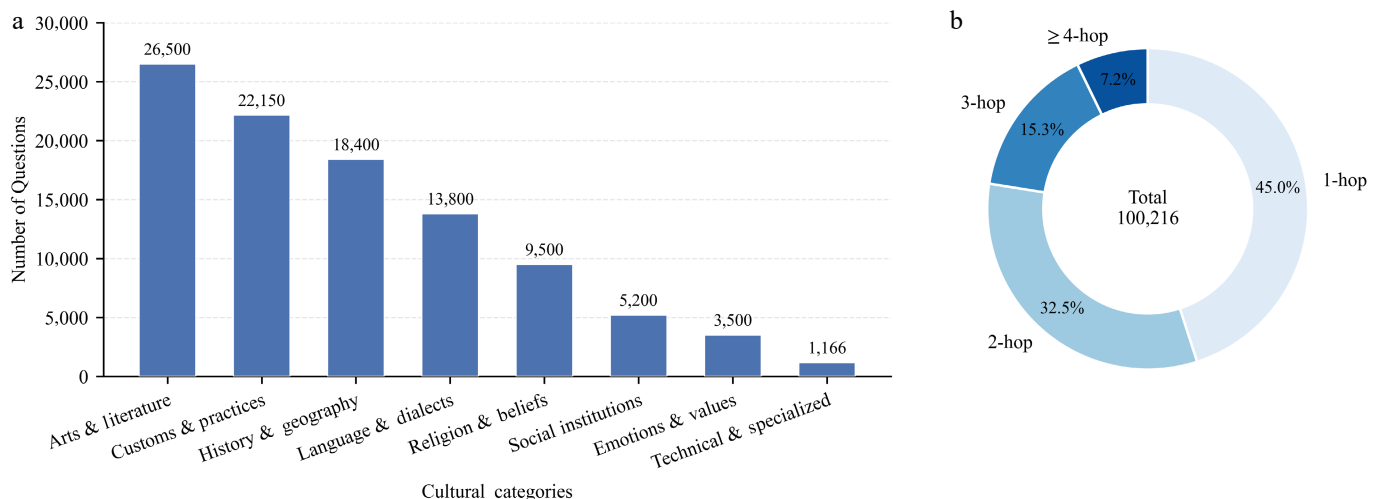


Fig. 4 Distribution statistics of the QA dataset. (a) Distribution of questions across cultural content categories; the horizontal axis represents the eight cultural categories, and the vertical axis represents the number of questions per category. (b) Distribution of questions by reasoning hops; the inner circle indicates the total number of questions, while the outer ring displays the percentages of 1-hop, 2-hop, 3-hop, and ≥ 4 -hop questions.

$$\text{Hits@1} = \frac{1}{|Q_{\text{test}}|} \sum_{q \in Q_{\text{test}}} \mathbb{I}(\text{rank}(q) = 1) \quad (4)$$

where Q_{test} denotes the test question set, $\mathbb{I}(\cdot)$ is the indicator function, and $\text{rank}(q)$ is the best rank of a predicted result that matches any gold answer for question q in the candidate answer list for that question.

• **Precision:** Precision measures the proportion of the model's predicted answers that are correct. Let the gold answer set for question q be $\mathcal{A}(q)$ and the predicted answer set be $\hat{\mathcal{A}}(q)$; then:

$$\text{Precision}(q) = \frac{|\hat{\mathcal{A}}(q) \cap \mathcal{A}(q)|}{|\hat{\mathcal{A}}(q)|}. \quad (5)$$

• **Recall:** Recall measures the proportion of gold answers that the model successfully retrieves. Accordingly, the recall for question q is defined as:

$$\text{Recall}(q) = \frac{|\hat{\mathcal{A}}(q) \cap \mathcal{A}(q)|}{|\mathcal{A}(q)|}. \quad (6)$$

• **F1-score:** The harmonic mean of precision (Precision) and recall (Recall), which jointly reflects the accuracy and coverage of the predicted answer set and is suitable for scenarios with multiple gold answers. In this work, the F1-score is computed from the overall Precision and Recall as follows:

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

5.3 Baselines

To thoroughly validate the performance of our framework, we compare it against seven representative baselines from two categories:

1. **LLM-only baselines:** These methods do not use an external knowledge graph and rely entirely on the LLM's internal knowledge to answer questions. They include **Zero-Shot LLM** and **Chain-of-Thought (CoT) LLM**^[39].

2. **KG-LLM baselines:** These are mainstream knowledge-graph question answering methods. They include **MindMap**^[38], **ToG**^[34], **PoG**^[35], **BYOKG-RAG**^[40], and **GraphRAG**^[41].

GraphRAG is an end-to-end pipeline that constructs a graph from raw text via LLM-based extraction and builds community-based summaries for query answering. Since our setting already provides the Chinese culture-loaded words knowledge graph, we skip GraphRAG's text-to-graph extraction and graph construction stages and only use its retrieval-related components, including community-level indexing and community-summary retrieval, on top of the given knowledge graph.

5.4 Implementation details

To validate the generality of our framework, we conduct experiments on three different large language models: GLM-4-Flash (<https://docs.bigmodel.cn/cn/guide/models/free/glm-4-flash-250414>), DeepSeek-V3 (<https://huggingface.co/deepseek-ai/DeepSeek-V3>), and Qwen-Plus (www.alibabacloud.com/help/en/model-studio/qwen-api-reference).

The auxiliary node-scoring dataset is separate from the end-to-end QA dataset. After deduplication over (question, node item, node label), it contains 5,137 unique question–node pairs. For the auxiliary scorer, we first split the data into a 60% development partition and a 40% held-out test partition. The development partition is split into training and validation subsets at an 8:2 ratio. The validation subset is used for early stopping and model selection, and the type-wise statistics reported in the Electronic Supplementary Information are computed only on the held-out test split.

For the node scoring model, we use the pretrained Chinese RoBERTa model `hfl/chinese-roberta-wwm-ext` (<https://huggingface>.

<https://huggingface> as the backbone encoder. The detailed hyperparameter settings are shown in Table 1. Training dynamics (training/validation loss curves) of the node scoring model are provided in Supplementary Fig. S4. Type-wise test statistics and their interpretation are provided in Supplementary Fig. S5, Supplementary Table S6, and Supplementary Note 3.

5.5 Results and analysis

We analyze the experimental results around two core research questions:

- **RQ1** How does the proposed method perform on the Chinese culture-loaded words question answering task?
- **RQ2** What factors primarily affect the overall performance on this task?

5.5.1 RQ1: Overall performance comparison

To answer RQ1, we conduct a comprehensive evaluation of our proposed method and all baseline methods on the constructed Chinese culture-loaded words question answering dataset. The experiments use three LLM backbones (GLM-4-Flash, DeepSeek-V3, and Qwen-Plus) and evaluate four metrics—Hits@1, Precision, Recall, and F1. The detailed results are reported in Table 2.

5.5.1.1 LLM-only vs KG-LLM augmentation methods

The LLM-only baselines (Zero-Shot and CoT) consistently lag behind all KG-LLM methods in terms of Hits@1, Precision, Recall, and F1 across the three LLM backbones. For instance, on DeepSeek-V3, CoT achieves only about 55% Hits@1, whereas CEAF-CLWQA raises Hits@1 to 87.1%, with Precision and Recall simultaneously reaching the 60%–70% range. These results indicate that, in knowledge-intensive and strongly context-dependent settings such as Chinese culture-loaded words question answering, relying solely on the LLM's internal parametric knowledge is insufficient to achieve adequate accuracy and coverage, whereas introducing a knowledge graph as externally grounded factual support can substantially improve answer quality.

5.5.1.2 Other KG-LLM methods vs our method

CEAF-CLWQA achieves the best Hits@1 across the three LLM backbones (76.3% for GLM-4-Flash, 87.1% for DeepSeek-V3, and 82.5% for Qwen-Plus), indicating that our method more reliably identifies the correct top-ranked answer. In terms of Precision/Recall/F1, on the two stronger backbones—DeepSeek-V3 and Qwen-Plus—our method outperforms the strongest KG-LLM baseline, PoG, across all three metrics; specifically, the F1 score improves by approximately 2.2 percentage points on DeepSeek-V3 and 0.6 percentage points on Qwen-Plus.

On GLM-4-Flash, our method still achieves the highest Hits@1 and Precision, but its Recall and F1 are lower than those of PoG.

Table 1. Hyperparameter settings for the node scoring model.

Category	Hyperparameter	Value
Model architecture	Backbone model	hfl/chinese-roberta-wwm-ext
	Node-type embedding dimension	64
	Fully connected layers	(768+64) → 256 → 64 → 1
Optimizer	Dropout rate	0.3
	Optimizer type	AdamW
	Learning rate	2e-5
	Weight decay	0.01
	Gradient clipping (max norm)	1.0
Training control	Batch size	16
	Max epochs	20
	Early-stopping patience	3

Table 2. Performance comparison of different methods on three LLM backbones.

Method type	Method	GLM-4-Flash				DeepSeek-V3				Qwen-Plus			
		Hits@1	Precision	Recall	F1	Hits@1	Precision	Recall	F1	Hits@1	Precision	Recall	F1
LLM-only	Zero-Shot	46.7	37.2	32.2	34.5	52.4	38.9	33.9	36.2	49.5	36.8	31.8	34.1
	CoT	52.1	44.1	39.2	41.5	54.8	41.8	36.8	39.1	53.9	44.1	39.1	41.4
KG-LLM	MindMap	67.9	48.9	53.9	51.3	76.2	60.0	65.0	62.4	74.4	56.1	61.1	58.5
	ToG	64.3	50.1	55.1	52.5	70.5	54.5	59.5	56.9	67.1	51.4	56.4	53.8
	BYOKG-RAG	69.5	53.2	52.8	53.0	80.2	61.5	66.2	63.8	76.5	59.8	64.5	62.1
	GraphRAG	70.1	57.8	54.1	55.9	84.6	63.1	68.2	65.6	79.4	62.2	67.1	64.5
	PoG	73.6	59.1	64.1	61.5	85.7	63.7	68.7	66.1	80.3	62.8	67.8	65.2
	CEAF-CLWQA (ours)	76.3	60.2	55.2	57.6	87.1	65.9	70.9	68.3	82.5	63.4	68.4	65.8

Note: Hits@1 denotes the proportion of questions for which at least one gold answer is ranked first; Precision denotes the proportion of predicted answers that are correct; Recall denotes the proportion of gold answers successfully retrieved; and F1 is the harmonic mean of Precision and Recall. All values are reported as percentages (%). Bold values indicate the best result for each metric among all compared methods under the same LLM backbone.

(55.2%/57.6% vs 64.1%/61.5%). This pattern suggests a backbone-dependent precision–coverage trade-off under weaker backbones: stricter cultural-element constraints may improve the reliability of the top-ranked answer, while answer completeness becomes more sensitive to the quality of early-stage constraints. To further investigate where this loss arises and how it may be mitigated, we provide a dedicated stage-wise robustness analysis in Section 5.5.4.

5.5.1.3 Comparison across different backbones

Across the three evaluated LLM backbones, our framework consistently improves Hits@1 over the LLM-only baselines by large margins, indicating that KG grounding is crucial for culture-loaded words question answering. Compared with the KG-LLM baselines in Table 2, CEAF-CLWQA achieves the best or near-best results in most settings and shows a consistent advantage in Hits@1 across all three backbones, while gains in answer completeness are more evident on the two stronger backbones. Taken together, these results suggest that the framework is effective across the evaluated backbones, but its benefits exhibit a backbone-dependent precision–coverage trade-off rather than uniform superiority in every metric.

5.5.1.4 Performance across reasoning hops

To further characterize reasoning difficulty, we report hop-wise results grouped by the length of the standard reasoning path (1-hop, 2-hop, 3-hop, and ≥ 4 -hop). Since PoG is the strongest baseline in Table 2, Table 3 compares PoG and CEAF-CLWQA under the same held-out test set and the same inference setting as the main experiment. The full hop-wise results for KG-LLM baselines are provided in Supplementary Tables S7–S9, with further analysis in Supplementary Note 4. Performance declines for both methods as the number of reasoning hops increases. On the two stronger backbones, however, CEAF-CLWQA shows a clearer advantage on 3-hop and ≥ 4 -hop questions than on 1-hop questions, especially in F1. This suggests that the progressive constraint mechanism contributes more substantially to complex multi-hop cultural reasoning than to simple one-step retrieval. A different pattern appears on GLM-4-Flash. CEAF-CLWQA still maintains higher Hits@1 across hop groups, indicating more reliable identification of the top-ranked core answer. However, its F1 falls behind PoG from deeper-hop questions onward, which is consistent with the stage-wise robustness analysis in Section 5.5.4.

5.5.2 RQ2: factor analysis

To answer RQ2, we investigate how key factors influence our framework's performance on the Chinese culture-loaded words question-answering task. Given that our method performs retrieval over the knowledge graph and aggregates multiple candidate evidence paths, we focus on two factors directly related to the candidate evidence

space: the number of retained cultural evidence paths, K , and the maximum search depth $D_{\max}$.

5.5.2.1 Effect of the number of retained cultural evidence paths K

Table 4 reports the F1 scores under different Top- K settings. The results show a steady increase in F1 as K gradually increases from 1 to 3. This indicates that, when a question has multiple gold answers, retaining an appropriate number of reasoning paths helps retrieve more culturally relevant evidence, thereby improving answer completeness. However, when K is further increased to 4 and 5, the score decreases, possibly because the additional paths introduce redundant or distracting information. Overall, retaining the Top-3 paths achieves a better balance between accuracy and answer completeness; therefore, we use $K = 3$ in our main experiments.

5.5.2.2 Effect of the maximum search depth

$D_{\max}$. Figure 5 illustrates the impact of the maximum search depth $D_{\max}$ on Hits@1 performance. We compare our full method with a variant that removes the path-ranking stage (w/o Ranking). The results show that when the search depth increases from 1 to 2, the performance of all methods improves substantially, demonstrating that culture-loaded words question answering generally relies on multi-hop reasoning paths. However, when the depth is further increased to 3, a critical divergence emerges across methods: as the number of candidate paths grows rapidly and introduces substantial noise, the variant without the ranking stage (w/o Ranking) suffers a cliff-like performance drop. In contrast, benefiting from the cultural-evidence-based

Table 3. Hop-wise comparison between PoG and CEAF-CLWQA on the three LLM backbones.

Backbone	Hop group	PoG Hits@1	PoG F1	CEAF Hits@1	CEAF F1
GLM-4-Flash	1-hop	79.9	65.1	83.5	65.1
	2-hop	73.9	62.3	76.7	56.8
	3-hop	64.4	56.2	65.4	47.2
	≥ 4 -hop	52.3	46.9	53.1	36.7
DeepSeek-V3	1-hop	89.4	69.4	90.0	70.0
	2-hop	86.7	67.3	88.2	69.7
	3-hop	79.9	60.5	82.0	64.8
	≥ 4 -hop	71.3	52.1	73.2	58.7
Qwen-Plus	1-hop	84.5	68.3	85.6	68.5
	2-hop	80.6	66.0	83.2	66.6
	3-hop	74.4	60.3	78.1	61.5
	≥ 4 -hop	64.7	52.7	69.1	54.8

Note: Hits@1 denotes the proportion of questions for which at least one gold answer is ranked first; F1 denotes the harmonic mean of Precision and Recall. All values are reported as percentages (%). The hop groups are defined according to the length of the standard reasoning path: 1-hop, 2-hop, 3-hop, and ≥ 4 -hop.

Table 4. F1 score (%) under different Top-K settings.

Model	Top-1	Top-2	Top-3	Top-4	Top-5
GLM-4-Flash	54.1	55.4	57.6	57.2	54.5
DeepSeek-V3	62.3	66.1	68.3	67.1	64.8
Qwen-Plus	61.8	62.4	65.8	63.3	60.4

Note: Top-K denotes the number of retained cultural evidence paths. All values are F1 scores reported as percentages (%). Bold values indicate the highest F1 score among the evaluated Top-K settings for each LLM backbone.

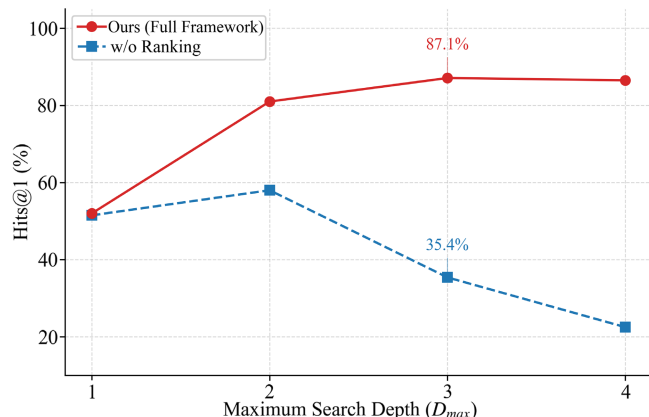


Fig. 5 Impact of maximum search depth (D_{max}) on Hits@1 performance. The figure illustrates the performance comparison between our full framework and the variant without ranking as the search depth increases. All results are obtained on DeepSeek-V3 with $K = 3$.

path-ranking mechanism, our full method can effectively suppress noise and precisely identify the core semantic paths, reaching peak performance at $D_{max} = 3$ (87.1%) and still maintaining a certain degree of robustness under the stress test of $D_{max} = 4$. Considering both computational efficiency and accuracy, we set the maximum search depth to 3 in the main experiments.

5.5.3 Ablation study

To systematically assess the practical contributions of key design choices and core stages in our method, we conduct ablation studies on DeepSeek-V3 and report the results in Table 5.

The results indicate that the four stages in our pipeline are highly coupled, and weakening any key component substantially degrades the overall performance. First, cultural-evidence-based path ranking is shown to be a crucial component for suppressing noise and improving answer quality: removing this stage reduces Hits@1 from 87.1% to 35.4%, with a marked decrease in F1 as well. This suggests that, after the preceding 'path retrieval and pruning guided by cultural semantic information', the candidate set still contains noisy paths that are structurally feasible but insufficiently context-relevant, and a ranking mechanism is necessary to distinguish and filter them effectively.

Second, cultural-element-based entity linking plays a guiding role in determining the retrieval starting point and constructing constraints.

Table 5. Ablation results (based on DeepSeek-V3).

Category	Method	Hits@1	F1
Full method	Our method	87.1%	68.3%
Module ablation	w/o cultural-element-based entity linking	72.1%	58.5%
	w/o cultural-evidence-based path ranking	35.4%	21.2%
Scoring-model ablation	w/o cultural-element embeddings	60.2%	43.7%
	Generic semantic-similarity ranking	55.8%	39.9%

Note: All values are reported as percentages (%). Bold values indicate the best performance among the full method and all ablation variants.

Without this stage, subsequent retrieval lacks the structured guidance provided by cultural-element constraints, making the candidate space more prone to expansion and the introduction of irrelevant evidence, which reduces Hits@1 to 72.1%; this validates that removing cultural-element constraints in entity linking directly weakens the effectiveness of subsequent path retrieval and evidence filtering.

Finally, within the scoring model design, cultural-element embeddings yield a clear discriminative gain: removing them reduces Hits@1 to 60.2%, which still outperforms a generic semantic-similarity ranking that relies only on textual semantics (55.8%).

This comparison further indicates that, in culture-loaded words question answering, cultural-element information provides a more stable discriminative signal when texts are similar but cultural elements differ, thereby improving the accuracy and consistency of evidence-path ranking.

5.5.4 Stage-wise robustness analysis across backbone strengths

To further diagnose the performance gap on GLM-4-Flash, we perform additional analyses on the same held-out test split used in Table 2, with $K = 3$ and $D_{max} = 3$. We focus on GLM-4-Flash and use DeepSeek-V3—the strongest backbone in our main experiments and the one used in the ablation study—as a stronger reference. For each question, we record the structured constraints produced in Stage 1, the candidate paths returned after Stage 2, the Top-3 paths retained after Stage 3, and the final predicted answers. We define an answer-supporting path as a candidate path that reaches at least one gold answer entity and is compatible with the question's standard reasoning path. Based on this definition, we examine five diagnostic indicators: (i) anchor-entity set F1 between the predicted and reference anchor-entity sets; (ii) element-label match, averaged over anchor-element labels and the target-element label; (iii) answer-supporting path coverage after retrieval; (iv) answer-supporting path retained in Top-3, computed only on the subset of questions already covered at Stage 2; and (v) complete-answer coverage on multi-answer questions, defined as the proportion of multi-answer questions for which all gold answers are retrieved. The reference anchor entities and element labels used in Tables 6 and 7 are derived from the annotated standard reasoning paths and ontology mappings used during dataset construction. When multiple element labels in T are ontologically compatible with the same gold path, we use the least specific label that remains consistent with the gold evidence path.

Table 6 shows that the larger gap between GLM-4-Flash and DeepSeek-V3 emerges mainly in Stage 1 and Stage 2. Compared with DeepSeek-V3, GLM-4-Flash obtains lower anchor-entity set F1 (88.4% vs 94.3%) and lower element-label match (84.1% vs 91.0%), which is followed by a 7.2-point drop in answer-supporting path coverage after retrieval (87.6% vs 94.8%). By contrast, the gap in conditional Top-3 retention after ranking is much smaller (91.6% vs 95.3%). These results indicate that the weaker backbone loses more answer coverage before the ranking stage, rather than because the ranking module itself becomes unstable. The substantially lower

Table 6. Stage-wise diagnosis across backbone strengths.

Diagnostic metric	GLM-4-Flash	DeepSeek-V3
Anchor-entity set F1	88.4%	94.3%
Element-label match	84.1%	91.0%
Answer-supporting path coverage after retrieval	87.6%	94.8%
Top-3 retention given Stage-2 coverage	91.6%	95.3%
Complete-answer coverage on multi-answer questions	44.8%	58.6%
Final Hits@1	76.3%	87.1%
Final F1	57.6%	68.3%

complete-answer coverage on multi-answer questions for GLM-4-Flash (44.8% vs. 58.6% for DeepSeek-V3) helps explain the observed performance pattern: GLM-4-Flash maintains a relatively strong Hits@1 score but achieves lower Recall and F1 scores.

To further localize the source of the degradation, we conduct controlled analyses on GLM-4-Flash by replacing the predicted Stage-1 constraints with gold Stage-1 constraints at inference time.

We also evaluate a lightweight constraint relaxation strategy to mitigate under-coverage caused by overly specific Stage-1 constraints under weaker backbones. To keep the fallback consistent with the main experimental setting, we use the same Top-K parameter as a soft target for the candidate budget. Since Table 4 shows that $K = 3$ provides the best balance between answer accuracy and answer completeness, the fallback is activated only when strict Stage-2 retrieval returns fewer than $K = 3$ deduplicated complete candidate paths.

More concretely, let C denote the candidate-path set produced by the default Stage-2 retrieval under the predicted anchor entities, anchor elements, and target element, with the same $D_{\max} = 3$ setting as in the main experiments. If $|C| \geq 3$, no relaxation is applied and the pipeline proceeds directly to the original ranking stage. If $|C| < 3$, we first relax only the target element by replacing it with its parent element in the ontology, while keeping the predicted anchor entities and anchor elements unchanged. Stage-2 retrieval is then rerun, and the newly retrieved paths are merged with C and deduplicated. If the resulting set still contains fewer than 3 candidate paths, we further relax anchor elements one by one from more specific to more general ontology levels. After each anchor-element backoff, Stage-2 retrieval is rerun, and the returned paths are again merged and deduplicated with the current candidate set.

Importantly, $K = 3$ is used here only as a target candidate budget for ranking, rather than as an assumption that every question should have at least three genuinely relevant paths. Therefore, the fallback stops once one of the following conditions is met: (i) the candidate set reaches size K ; (ii) no further parent element is available for relaxation; or (iii) the latest relaxation step introduces no new deduplicated path. Finally, the resulting candidate set is passed to the same path-ranking module used in the default pipeline. If more than 3 candidate paths are available, we retain the Top-3 ranked paths; otherwise, all remaining paths are kept for answer generation.

Table 7 shows that both gold anchor entities and gold cultural elements substantially recover Recall/F1, and the full Stage-1 control yields the largest gain (+7.8 Recall and +4.8 F1). This indicates that the weaker-backbone gap is explained primarily by the quality of Stage-1 constraint extraction rather than by instability in the ranking stage. The relaxation variant also improves Recall from 55.2% to 58.6% and F1 from 57.6% to 59.2%, while decreasing Hits@1 only slightly (76.3% to 75.9%). This suggests that moderate ontology-aware backoff can partially recover answer completeness under weaker backbones without fundamentally changing the framework.

Taken together, these results support a backbone-dependent empirical observation rather than a universal capability floor. Within the evaluated backbones, CEAF-CLWQA is strictly superior to PoG on both Hits@1 and F1 for the stronger backbone considered in the

Table 7. Controlled analyses of Stage-1 constraints and a lightweight constraint relaxation strategy on GLM-4-Flash.

Setting	Hits@1	Precision	Recall	F1
CEAF-CLWQA (original)	76.3%	60.2%	55.2%	57.6%
+ Gold anchor entities	79.4%	60.9%	60.1%	60.5%
+ Gold cultural elements	78.6%	60.6%	59.1%	59.8%
+ Gold full Stage-1 constraints	81.1%	61.8%	63.0%	62.4%
+ constraint relaxation	75.9%	59.8%	58.6%	59.2%

stage-wise analysis (DeepSeek-V3), whereas on GLM-4-Flash it shows a precision–coverage trade-off. Because the tested backbones are heterogeneous in architecture, training data, and serving configuration, we do not infer a universal threshold in parameter count or benchmark score. Instead, the practically relevant floor in our setting is whether Stage-1 constraint extraction is sufficiently stable to preserve answer-supporting path coverage before ranking.

5.5.5 Computational efficiency analysis

In addition to effectiveness, we further evaluate the computational efficiency of CEAF-CLWQA by comparing its inference cost with representative baselines. All methods were evaluated under the same hardware and network environment. The reported latency is the average end-to-end inference time per query.

As shown in Table 8, Zero-Shot is the most efficient baseline in terms of inference latency, since it directly generates an answer with only one LLM call per query. In contrast, ToG requires substantially more LLM interactions on average, leading to the highest latency among the compared methods. Although CEAF-CLWQA adopts a multi-stage pipeline, its online interaction with the LLM remains fixed at two calls per query, namely cultural-element-based entity linking and final answer generation. As a result, CEAF-CLWQA achieves lower latency than ToG, while maintaining a fixed two-call LLM interaction pattern. This suggests that the additional stages in CEAF-CLWQA mainly introduce bounded symbolic retrieval and ranking overhead rather than repeated LLM reasoning.

Table 9 further presents the component-wise latency breakdown of CEAF-CLWQA. Among all stages, cultural-element-based entity linking and answer generation together account for 55.3% of the total inference time, indicating that the LLM-involved stages remain the dominant source of latency. By comparison, path retrieval and pruning contribute only 11.2% of the total latency, suggesting that constrained graph retrieval itself is relatively efficient. The remaining non-LLM overhead is mainly concentrated in the ranking stage, which accounts for 33.5% of the total latency. Overall, these results show that the computational cost of CEAF-CLWQA is not caused by iterative LLM calls, but by a bounded combination of LLM-based understanding, constrained graph operations, and evidence-aware ranking.

6 Discussion

6.1 Implications for KG-enhanced LLM methods

Our results suggest that, for culture-loaded words question answering, the value of KG enhancement lies not in retrieving more paths, but in converting question-relevant cultural cues into explicit

Table 8. Computational efficiency comparison of CEAF-CLWQA and baselines.

Method	Avg. LLM calls/query	Avg. latency (s/query)
Zero-Shot	1	5.50
ToG	7.8	32.92
CEAF-CLWQA (ours)	2	29.14

Table 9. Component-wise inference latency breakdown of CEAF-CLWQA.

Stage	LLM involved?	Avg. latency (s/query)	Share
Cultural-element-based entity linking	Yes	7.62	26.1%
Path retrieval and pruning	No	3.24	11.2%
Node scoring and path ranking	No	9.76	33.5%
Answer generation	Yes	8.52	29.2%
Total	—	29.14	100.0%

constraints and applying them consistently across entity linking, retrieval/pruning, and ranking. Cultural-element constraints reduce retrieval drift early by narrowing the search space, while cultural-evidence-based ranking becomes more important once multi-hop retrieval produces many structurally plausible but contextually mismatched candidate paths. The observed Top-K trade-off further shows that answer completeness does not increase monotonically with more retrieved evidence, because excessive candidates mainly introduce noise.

At the same time, the present results do not support a universal capability floor expressed by parameter count or a single benchmark score. Across the three evaluated backbones, CEAF-CLWQA is consistently superior to PoG in Hits@1, but superior in F1 only on DeepSeek-V3 and Qwen-Plus, not on GLM-4-Flash. The stage-wise analyses indicate that this weaker-backbone gap mainly originates before ranking, especially in Stage-1 constraint extraction and answer-supporting path coverage. Thus, the more informative operational condition is not model scale itself, but whether Stage-1 extraction remains sufficiently stable. When it does, progressive cultural-element constraints improve both top-1 reliability and answer completeness; when it does not, the framework exhibits a precision-coverage trade-off.

6.2 Implications for cultural question answering research and evaluation

For cultural question answering, the central challenge is not only answer generation, but evidence-grounded contextual interpretation. Questions about culture-loaded words often depend on provenance, allusive background, usage context, and interpretive scope. If such dimensions remain only implicit in the prompt, they can be weakened during multi-hop retrieval or replaced by textually similar but contextually mismatched evidence. Our results therefore support representing cultural cues as structured constraints rather than leaving them at the prompt level only.

The evaluation implications are equally important. Aggregate end-answer metrics remain necessary, but they are not sufficient for this task. The hop-wise results show that the benefit of CEAF-CLWQA is not uniform across reasoning difficulty: on the two stronger backbones, its gains become clearer on deeper-hop questions, whereas on GLM-4-Flash the method mainly preserves stronger Top-1 reliability while losing answer completeness on deeper questions. More generally, the stage-wise diagnostics show that similar final answer scores may conceal different failure modes, including unstable constraint extraction, insufficient retrieval coverage, or errors in evidence selection. For this reason, future work on cultural question answering should report not only aggregate accuracy, recall, and F1, but also hop-stratified results and evidence-oriented diagnostics such as evidence relevance, evidence coverage, and citation consistency.

6.3 Practical implications and application boundaries

In practical settings, CEAF-CLWQA is better suited to scenarios where answers need to be accompanied by provenance and interpretable evidence, such as cultural education and digital-humanities knowledge services. From an engineering perspective, its online interaction pattern remains predictable: the framework uses two LLM calls per query, and the additional overhead mainly comes from bounded symbolic retrieval and ranking rather than repeated LLM reasoning.

However, the scope and transfer conditions of the cultural-element inventory used in this work should be stated explicitly. This inventory is not intended as a universal culture-independent set; it is derived from the ontology of the current Chinese culture-loaded words knowledge graph and is therefore domain-dependent. Accordingly,

transferring CEAF-CLWQA to another cultural setting would not require rebuilding the full framework from scratch, but it would require adapting the target ontology and its associated cultural-element inventory. The transferable part is the pipeline itself—constraint extraction, ontology-guided retrieval and pruning, evidence ranking, and evidence-constrained answer generation—whereas the culture-specific part is the inventory of fine-grained cultural elements and their ontology mappings. Because the ontology is hierarchical, this adaptation can proceed in a coarse-to-fine manner: reusable upper-level evidential dimensions, such as provenance, source type, person, time period, and quotation/explanation context, can first be aligned across cultures; lower-level culture-specific elements can then be bootstrapped from target-domain corpora and existing ontologies/knowledge graphs, and finally normalized through expert validation. This transfer should not be interpreted as zero-shot reuse of all learned components: because the node scoring model learns cultural-element embeddings over the ontology-derived label inventory and is trained on question-node supervision from the current KG, cross-cultural deployment would ordinarily require retraining or fine-tuning this ranking component on target-domain data, even if the overall pipeline is retained. This coarse-to-fine view is also consistent with the ontology-aware backoff mechanism used in our robustness analysis, where constraints can be relaxed from more specific to more general ontology levels when strict fine-grained labels lead to under-coverage. Accordingly, the present results should be interpreted as empirical findings under the current dataset, ontology, and evaluated backbones.

7 Conclusions

Focusing on Chinese culture-loaded words question answering—a knowledge-intensive and strongly context-dependent setting—we propose a cultural-element-augmented KG-LLM question answering framework to address several key challenges: LLMs are prone to factual errors and hallucinated content when domain knowledge coverage is insufficient, and existing generic KG-LLM methods suffer from the lack of cultural-element constraints in entity linking, the absence of cultural semantic information guidance in path retrieval and pruning, and inaccurate ranking of cultural evidence paths in this setting. Our framework explicitly injects constraints from cultural elements, cultural semantic information, and a cultural evidence selection mechanism into key decision stages. It employs a hierarchical constraint process—"cultural-element-based entity linking—path retrieval and pruning guided by cultural semantic information—cultural-evidence-based path ranking"—to reduce erroneous retrieval propagation caused by unstable entity alignment, suppress the accumulation of irrelevant path noise, and decrease reliance on the LLM's subjective judgments at critical steps. In the implementation, the framework sequentially performs cultural-element constraint extraction and entity linking and retrieves candidate paths subject to cultural elements and structural constraints. It prunes them using cultural semantic information, ranks candidate paths by cultural-element signals to consistently select the Top-K evidence paths, and finally generates a more comprehensive natural-language answer based on the Top-K cultural evidence paths. Experiments on three LLM backbones validate the effectiveness of the proposed framework: CEAF-CLWQA achieves the best Hits@1 in all settings and outperforms strong baselines such as ToG and PoG on F1 under the two stronger backbones. Additional stage-wise analyses further show that, on the weaker backbone, the framework exhibits a precision-coverage trade-off: it remains more reliable in identifying the top-ranked core answer, while answer completeness is more sensitive to the quality of Stage-1 constraint extraction. Moreover, we

construct a Chinese culture-loaded words question answering dataset containing 100,216 question-answer pairs, providing a reusable technical resource for knowledge-augmented question-answering research in the cultural domain.

Author contributions

The authors confirm their contributions to this study as follows: conceptualization: Xu F, Gu J; methodology, software: Xu F, Zhang W; data curation: Xu F, Tu W, Liu H; Resources: Zhang W, Tu W; investigation: Xu F; writing - original draft preparation: Xu F; validation: Zhang W, Liu H; writing - review & editing: Zhang W, Tu W, Liu H, Xiao Q, Gu J; supervision: Xiao Q; funding acquisition: Xiao Q, Gu J. All authors reviewed the results and approved the final version of the manuscript.

Data availability

We have publicly released the materials most directly related to this study at the GitHub repository <https://github.com/qwzwdaf/CLWQA>.

The repository includes: (i) the 12 QA templates used for dataset construction; (ii) a sampled subset of the QA data; (iii) a sampled subset of the auxiliary node-scoring data; and (iv) sample shards of the Chinese culture-loaded words knowledge graph.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFC3300801), the Research Project of China Publishing Promotion Association (Grant No. 2025ZBCH-JYYB17), and the Natural Science Foundation of Hubei Province (CN) (Grant No. 2025AFB078).

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper online at: <https://doi.org/10.48130/ker-0026-0011>.

Dates

Received 19 January 2026; Revised 24 March 2026; Accepted 7 May 2026; Published online 28 September 2026

References

- [1] Nida EA, Taber CR. 1982. *The theory and practice of translation: helps for translators*. 2nd Edition. Leiden: E.J. Brill.
- [2] Newmark P. 1988. *A textbook of translation*. New York: Prentice-Hall International.
- [3] Nida EA. 1993. *Language, culture, and translating*. Shanghai: Shanghai Foreign Language Education Press
- [4] Chang Y, Wang X, Wang J, Wu Y, Yang L, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15(3):39
- [5] Wang L, Ma C, Feng X, Zhang Z, Yang H, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18(6):186345
- [6] Zhao Z, Fan W, Li J, Liu Y, Mei X, et al. 2024. Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering* 36(11):6889–6907
- [7] Huang L, Yu W, Ma W, Zhong W, Feng Z, et al. 2025. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2):1–55
- [8] Li D, Sun Z, Hu X, Liu Z, Chen Z, et al. 2023. A survey of large language models attribution. *arXiv Preprint* 2311.03731
- [9] Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy SMTI, et al. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, December 2023. USA: Association for Computational Linguistics. pp. 2541–2573 doi: [10.18653/v1/2023.emnlp-main.155](https://doi.org/10.18653/v1/2023.emnlp-main.155)
- [10] Luo Y, Liu Y, Zhang L, Gao F, Gu J. 2025. A survey on quality evaluation of instruction fine-tuning datasets for large language models. *Data Intelligence* 7(3):527–566
- [11] Kim J, Kwon Y, Jo Y, Choi E. 2023. KG-GPT: a general framework for reasoning on knowledge graphs using large language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore, December 2023. USA: Association for Computational Linguistics. pp. 9410–9421 doi: [10.18653/v1/2023.findings-emnlp.631](https://doi.org/10.18653/v1/2023.findings-emnlp.631)
- [12] Wei Y, Huang Q, Zhang Y, Kwok J. 2023. KICGPT: large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore, December 2023. USA: Association for Computational Linguistics. pp. 8667–8683 doi: [10.18653/v1/2023.findings-emnlp.580](https://doi.org/10.18653/v1/2023.findings-emnlp.580)
- [13] Yang L, Chen H, Li Z, Ding X, Wu X. 2024. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering* 36(7):3091–3110
- [14] Pan S, Luo L, Wang Y, Chen C, Wang J, et al. 2024. Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36(7):3580–3599
- [15] Herscovitch D, Frank S, Lent H, de Lhoneux M, Abdou M, et al. 2022. Challenges and strategies in cross-cultural NLP. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland, May 2022. USA: Association for Computational Linguistics. pp. 6997–7013 doi: [10.18653/v1/2022.acl-long.482](https://doi.org/10.18653/v1/2022.acl-long.482)
- [16] Liu H, Cao Y, Wu X, Qiu C, Gu J, et al. 2025. Towards realistic evaluation of cultural value alignment in large language models: diversity enhancement for survey response simulation. *Information Processing & Management* 62(4):104099
- [17] Liu H, Li Q, Gao C, Cao Y, Xu X, et al. 2025. Beyond demographics: Enhancing cultural value survey simulation with multi-stage personality-driven cognitive reasoning. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China, November 2025. USA: Association for Computational Linguistics. pp. 18406–18428 doi: [10.18653/v1/2025.emnlp-main.928](https://doi.org/10.18653/v1/2025.emnlp-main.928)
- [18] Pawar S, Park J, Jin J, Arora A, Myung J, et al. 2025. Survey of cultural awareness in language models: text and beyond. *Computational Linguistics* 51(3):907–1004
- [19] Chiu YY, Jiang L, Lin BY, Park CY, Li SS, et al. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria, July 2025. USA: Association for Computational Linguistics. pp. 25663–25701 doi: [10.18653/v1/2025.acl-long.1247](https://doi.org/10.18653/v1/2025.acl-long.1247)
- [20] Arora S, Karpinska M, Chen HT, Bhattacharjee I, Iyyer M, et al. 2025. CaLMQA: exploring culturally specific long-form question answering across 23 languages. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria, July 2025. USA: Association for Computational Linguistics. pp. 11772–11817 doi: [10.18653/v1/2025.acl-long.578](https://doi.org/10.18653/v1/2025.acl-long.578)
- [21] Etxaniz J, Azkune G, Soroa A, De Lacalle O, Artetxe M. 2024. BertaQA: how much do language models know about local culture? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track, 10–15 December 2024, Vancouver, Canada*. USA:

- Neural Information Processing Systems Foundation, Inc. (NeurIPS). pp. 34077–34097 doi: [10.52202/079017-1073](https://doi.org/10.52202/079017-1073)
- [22] Mousi B, Durrani N, Ahmad F, Hasan MA, Hasanain M, et al. 2025. AraDiCE: benchmarks for dialectal and cultural capabilities in LLMs. *Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE, January 2025*. USA: Association for Computational Linguistics. pp. 4186–4218 <https://aclanthology.org/2025.coling-main.283>
- [23] Sun J, Huang W, Wu J, Gu C, Li W, et al. 2024. Benchmarking Chinese commonsense reasoning of LLMs: from Chinese-specifics to reasoning-memorization correlations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, August 2024*. USA: Association for Computational Linguistics. pp. 11205–11228 doi: [10.18653/v1/2024.acl-long.604](https://doi.org/10.18653/v1/2024.acl-long.604)
- [24] Cheng Q, Sun T, Zhang W, Wang S, Liu X, et al. 2023. Evaluating hallucinations in Chinese large language models. *arXiv Preprint 2310.03368*
- [25] Li Y, Zhang G, Qu X, Li J, Li Z, et al. 2024. CIF-bench: a Chinese instruction-following benchmark for evaluating the generalizability of large language models. *Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, August 2024*. USA: Association for Computational Linguistics. pp. 12431–12446 doi: [10.18653/v1/2024.findings-acl.739](https://doi.org/10.18653/v1/2024.findings-acl.739)
- [26] Doerr M. 2003. The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata. *AI Magazine* 24(3):75–92
- [27] Carriero VA, Gangemi A, Mancinelli ML, Marinucci L, Nuzzolese AG, et al. 2019. ArCo: the Italian cultural heritage knowledge graph. *The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science*, ed. Ghidini C. Cham: Springer. pp. 36–52 doi: [10.1007/978-3-030-30796-7_3](https://doi.org/10.1007/978-3-030-30796-7_3)
- [28] Koho M, Ikkala E, Leskinen P, Tamper M, Tuominen J, et al. 2021. WarSampo knowledge graph: Finland in the Second World War as linked open data. *Semantic Web* 12(2):265–278
- [29] Xu L, Lu L, Liu M, Song C, Wu L. 2024. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Heritage Science* 12:118
- [30] Yuan H, Li Y, Wang B, Liu K, Zhang J. 2025. Knowledge graph-based intelligent question answering system for ancient Chinese costume heritage. *npj Heritage Science* 13:198
- [31] Zhang W, Xiao QL, Liu HJ, Ren H, Cai ZY, et al. 2025. Large model-assisted extraction and knowledge graph construction of Chinese culture-loaded words. *Digital Library Forum* 21(1):33–45
- [32] Ma C, Chen Y, Wu T, Khan A, Wang H. 2025. Large language models meet knowledge graphs for question answering: synthesis and opportunities. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China, November 2025*. USA: Association for Computational Linguistics. pp. 24578–24597 doi: [10.18653/v1/2025.emnlp-main.1249](https://doi.org/10.18653/v1/2025.emnlp-main.1249)
- [33] Baek J, Aji AF, Saffari A. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE). Toronto, Canada, June 2023*. USA: Association for Computational Linguistics. pp. 78–106 doi: [10.18653/v1/2023.nlrse-1.7](https://doi.org/10.18653/v1/2023.nlrse-1.7)
- [34] Sun J, Xu C, Tang L, Wang S, Lin C, et al. 2024. Think-on-Graph: deep and responsible reasoning of large language model on knowledge graph. *International Conference on Learning Representations 2024 (ICLR 2024), Vienna, Austria, 7–11 May 2024*. Appleton, WI, USA: ICLR. pp. 3868–3898 https://proceedings.iclr.cc/paper_files/paper/2024/hash/10a6bdcabbd5a3d36b760daa295f63c1-Abstract-Conference.html
- [35] Tan X, Wang X, Liu Q, Xu X, Yuan X, et al. 2025. Paths-over-Graph: knowledge graph empowered large language model reasoning. *WWW '25: Proceedings of the ACM on Web Conference 2025. Sydney NSW, Australia, 2025*. New York, USA: Association for Computing Machinery. pp. 3505–3522 doi: [10.1145/3696410.3714892](https://doi.org/10.1145/3696410.3714892)
- [36] Jiang J, Zhou K, Dong Z, Ye K, Zhao X, et al. 2023. StructGPT: a general framework for large language model to reason over structured data. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023*. USA: Association for Computational Linguistics. pp. 9237–9251 doi: [10.18653/v1/2023.emnlp-main.574](https://doi.org/10.18653/v1/2023.emnlp-main.574)
- [37] Luo L, Li YF, Haffari R, Pan S. 2024. Reasoning on graphs: faithful and interpretable large language model reasoning. *International Conference on Learning Representations 2024 (ICLR 2024), Vienna, Austria, 7–11 May 2024*. Appleton, WI, USA: ICLR. pp. 14400–14423 https://proceedings.iclr.cc/paper_files/paper/2024/file/3e2aeb66481dd63a32421bf032b70384-Paper-Conference.pdf
- [38] Wen Y, Wang Z, Sun J. 2024. MindMap: knowledge graph prompting sparks graph of thoughts in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, August 2024*. USA: Association for Computational Linguistics. pp. 10370–10388 doi: [10.18653/v1/2024.acl-long.558](https://doi.org/10.18653/v1/2024.acl-long.558)
- [39] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, Louisiana, USA, 28 November – 9 December 2022*. San Diego, CA, USA: NeurIPS. pp. 24824–24837 https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- [40] Mavromatis C, Adeshina S, Ioannidis VN, Han Z, Zhu Q, et al. 2025. BYOKG-RAG: multi-strategy graph retrieval for knowledge graph question answering. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China, November 2025*. USA: Association for Computational Linguistics. pp. 27881–27898 doi: [10.18653/v1/2025.emnlp-main.1417](https://doi.org/10.18653/v1/2025.emnlp-main.1417)
- [41] Edge D, Trinh H, Cheng N, Bradley J, Chao A, et al. 2025. From local to global: a graph RAG approach to query-focused summarization. *arXiv Preprint 2404.16130*



Copyright: © 2026 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.