Prompting Large Language Models for Automatic Question Tagging

Nuojia Xu[1,2], Dizhan Xue[1,2], Shengsheng Qian[1,2], Quan Fang[3] and Jun Hu[4]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.
[3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.
[4]School of Computing, National University of Singapore, Beijing 119077, Singapore.

Abstract

Automatic question tagging (AQT) represents a crucial task in Community Question Answering (CQA) websites. Its pivotal role lies in substantially augmenting user experience through the optimization of question-answering efficiency. Existing question tagging models focus on the features of questions and tags, ignoring the external knowledge of the real world. Large language models can work as knowledge engines for incorporating real-world facts for different tasks. However, it is difficult for large language models to output tags in the database of CQA websites. To address this challenge, we propose a Large Language Model Enhanced Question Tagging method called LLMEQT to perform the question tagging task. In LLMEQT, a traditional question tagging method is first applied to pre-retrieve tags for questions. Then prompts are formulated for LLMs to comprehend the task and select more suitable tags from the candidate tags for questions. Results of our experiments on two real-world datasets demonstrate that LLMEQT significantly enhances the automatic question tagging performance for CQA, surpassing the performance of state-of-the-art methods.

Keywords: Community Question Answering, Machine Learning, Large Language Model, Prompt Learning, Question Tagging

1 Introduction

Community Question Answering (CQA) websites, for instance, Quora and Zhihu, serve as reliable repositories where users tend to access valuable information through posted questions and their corresponding answers [1] [2] [3]. Within these websites, questions are associated with diverse tags that encapsulate the underlying themes of questions, aiding users in locating potential information. The integration of these tags significantly augments the efficacy of various applications derived from CQA websites. These applications cover multiple fields, encompassing recommendation systems [4, 5], expert-finding systems [6–8], and search engines [9, 10]. However, a significant challenge prevalent in CQA websites pertains to the issue of incomplete assignment of tags to questions by users. This limitation compromises the optimal functionality of these websites.

Given the important role of tags in these applications, Automatic Question Tagging (AQT) has gained attention. Current AQT [11, 12] approaches primarily focus on analyzing textual information and inherent structural attributes within questions and their associated tags. Despite their demonstrated success, these approaches are constrained by their limited consideration of external real-world knowledge. The integration of such external knowledge holds immense promise in substantially enhancing the effectiveness of Automatic Question Tagging (AQT). Introducing external knowledge to AQT facilitates the inclusion of crucial factual information from external sources. It aims to deepen the comprehension of both questions and tags by incorporating real-world facts, thereby enriching the understanding of the inherent semantic nuances within the content.

A direct and effective method for knowledge-based approaches involves the utilization of pre-trained large language models (LLMs), e.g., GPT-3 [13], which have been trained on massive real-world knowledge. These large language models possess universal capabilities and demonstrate capability across a spectrum of tasks. To utilize LLMs effectively, designing appropriate prompting strategies is of value. Clearly describing the task goal and presenting proper examples from the training data significantly enhances the comprehensibility of instructions for LLMs. In the question tagging task, leveraging LLMs allows for a comprehensive analysis of the underlying context within questions, incorporating external knowledge to aid in accurate tagging. The concept of Automatic Question Tagging (AQT) enhanced by LLMs is summarized in Figure 1. Consider the question, 'Is our diet much better than that of Zhou Tianzi?'. Here, 'Zhou Tianzi' refers to the emperors of the Zhou Dynasty. Without external knowledge, understanding such proper nouns becomes challenging, causing the question tagging model to primarily rely on

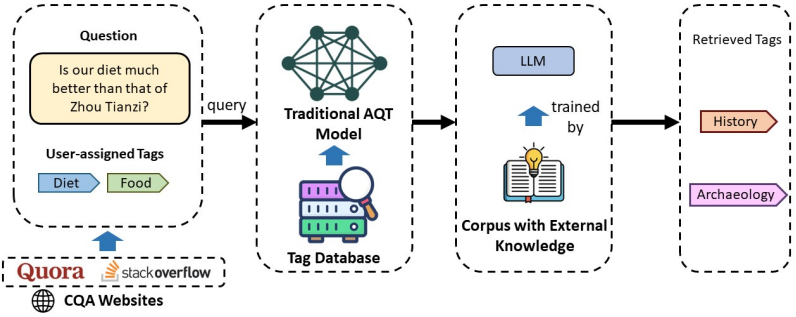


Fig. 1 The proposed workflow for AQT with LLM. The question tagging model help retrieve more tags for the question in CQA websites. Then the LLM, which is trained by corpus with external knowledge, will help understand and select the proper tags from the candidate tags.

noun embeddings, which depend on the impact of the encoders. Leveraging external knowledge via LLMs enhances the model's ability to comprehend the deep meaning of questions and associated tags, therefore improving the AQT task.

Despite the considerable efficacy of LLMs in incorporating external knowledge, there exist several limitations inherent in the AQT task. Notably, the tags generated by LLMs may not align with the existing database of CQA websites. Users, when attempting to append tags to questions, are constrained to selecting from pre-existing tags within the database rather than allocating much time to create new ones. For instance, as illustrated in Figure 1, an LLM might suggest tags such as 'King Wen of Zhou' or 'King Wu of Zhou' for a question like 'Is our diet much better than that of Zhou Tianzi?'. Nevertheless, these suggested tags may not exist within the database of the CQA websites. Additionally, newly created tags within CQA websites may lack contextual coherence and fail to facilitate the effective presentation of questions to users. Therefore, we should address **Challenge 1**: how to appropriately tag the questions with tags that exist in CQA websites by LLM?

Furthermore, LLMs operate within a few-shot learning paradigm that requires a small number of in-context examples to effectively adapt to diverse tasks. In the context of the question tagging task, the instructions provided to LLMs hold considerable importance as these LLMs generate outputs based on the given prompts. The selected examples of questions and associated tags significantly influence the model's performance. Improper examples can lead to sub-optimal outputs from LLMs and the generated content will be unusable. Thus, we need to address **Challenge 2**: how to design proper prompts for LLMs?

To tackle the above challenges, we propose a Large Language Model Enhanced Question Tagging method called LLMEQT, which is designed for question tagging enhanced by LLM. To address **Challenge 1**, we use the traditional question tagging model to pre-retrieve tags for questions and then

reserve them. By utilizing features of tags and questions, several tags can summarize the topic of the question in the first stage. To address **Challenge 2**, we design the prompts for LLMs to select tags from candidate tags. LLMs will integrate external knowledge to comprehend the deep meaning of questions and tags, thereby enhancing the automatic question tagging task. To evaluate our model, thorough experiments are taken. Our proposed model exhibits superiority over state-of-the-art methods in performance when compared across two real-world datasets.

In this paper, our primary contributions can be summarized in threefold:

- We designed a Large Language Model Enhanced Question Tagging method called LLMEQT tailored for the AQT task in CQA websites. It combines traditional question tagging models with the advanced capabilities of LLMs to perform question tagging. This approach leverages both the inherent features and the semantics in the original text of both questions and tags to enhance the effectiveness of the task.
- We formulate prompts for LLMs to enable a comprehensive understanding of the task, and facilitate the generation of accurate results. This process involves a strategic combination of instructions and proper instances, allowing LLMs to select appropriate tags for questions.
- We conduct experiments on two real-world datasets, namely Zhihu and Zhuanzhi, to validate the effectiveness of our model. These datasets are sourced from authentic CQA websites. The comprehensive experiments showcase the efficacy of our model in comparison to state-of-the-art methods in the realm of AQT.

2 Related Work

2.1 Question Tagging

In CQA websites, AQT holds an important role in diverse functions, including recommendation systems [4, 5], expert-finding systems [6, 7], and search engines [9, 14, 15]. In recommendation systems, tags play a crucial role in extracting essential question information, and they facilitate the provision of appropriate answers to user queries [16, 17]. In expert-finding systems, the utilization of question-specific tags empowers users to evaluate the alignment of retrieved information with their specific requirements [18, 19]. Furthermore, in search engines, leveraging tags can enhance the capability to pinpoint relevant information, helping users locate the information they seek [10, 20].

Several works have focused on AQT to integrate supplementary information for CQA websites. Initially, researchers focused on the classification of minority tags. Liu et al. [21] clustered questions with similar tags to highlight the key information of questions, and overcoming the problem of over-generalization of tags. Wasim et al. [22] generated a multi-labeled corpus by exploring the process of Question Answering system. By exploiting the dependence between tags of a particular question, the method worked in the task of biomedical

question classification. Moreover, researchers also worked on domain-agnostic class tags [23] to model questions by phrase-based tags. The proposed approach incorporated a new tag regularization mechanism for mapping questions to class tags.

Recently, researchers have shown a growing interest in focusing on question tagging tasks rather than the question classification tasks. Nie et al. [24] constructed a Directed Acyclic Graph (DAG) for tags in CQA websites. The features can be transmitted by the hierarchical relations. And the model retrieved tags by the embedding interaction. The utilization of a hierarchical learning taxonomy [12] enables the search for questions based on their respective tags. Additionally, Zhang et al. [25] concentrated on tag relations in CQA websites. Their model incorporated message passing from parent tag nodes to child tag nodes, allowing the model to effectively tag questions even when the tags are unseen.

Different from the aforementioned methodologies, our proposed question-tagging approach incorporates real-world facts to enhance the efficacy of the task.

2.2 Prompting for LLM

Recently, Large Language Models have shown promising abilities to integrate external knowledge for different tasks. The major approach to utilizing LLMs is prompting [26]. By designing proper prompts, LLMs could adapt to diverse tasks. In-context learning (ICL) [13, 27] is a popular method for utilizing LLMs. It first designed task descriptions as demonstrations for specific tasks, then selected a few examples from the task datasets [28–30]. Also, some approaches evaluate the example set as a whole to choose the most representative set of examples for specific tasks [31]. When selecting demonstrations, it is also necessary to take the relevance and diversity of examples into consideration [32]. Chain-of-Thought (CoT) [33] prompting method is an improved approach of prompting learning. In addition to demonstration examples of input and output, intermediate reasoning steps are also added to prompts to assist large language models in executing more complex tasks. Self-consistency [34] is a sampling-based method. It generates different reasoning paths, then finds consistency from all results and selects the most stable answer. Ling et al. [35] designed a special prompting format to make the LLMs self-verify step-by-step to confirm the correctness of reasoning steps. Kojima et al. [36] generated inference steps without relying on input and output examples. This method guides LLMs to perform step-wise inference, enabling them to achieve ideal results when dealing with complex inference tasks.

In CQA websites, the questions cover various fields and require external knowledge for better understanding. The question-answering tasks benefit a lot from LLMs. Shao et al. [37] prompted GPT-3 with answer heuristics for knowledge-based visual question answering (VQA). By extracting two types of complementary answer heuristics from a vanilla model, researchers encoded

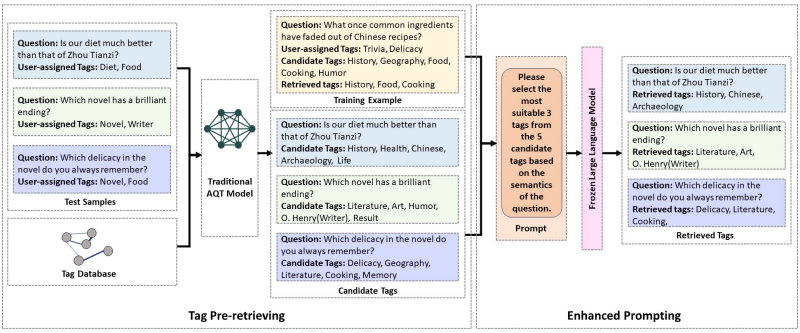


Fig. 2 The framework of LLMEQT: (1) We first use the traditional automatic question tagging model to pre-retrieve tags from the tag database for the questions. (2) We formulate prompts by combining instructions and examples to instruct the large language model. Then the large language model selects the more suitable tags from the candidate tags for questions.

them into the prompts for GPT-3 to better understand the task and give better answers. A plug-and-play module Img2LLM [38] was proposed to provide LLMs with special prompts that enable LLMs to perform zero-shot VQA tasks. What is more, the method is without end-to-end training, reducing calculating cost. Kim et al. [39] proposed a novel framework which is called Flipped-VQA. It encourages the model to predict all the combinations of visual question-answering triplets. By flipping the source pair and the target label, it can deal with complex relations.

These methods confirm that prompts for LLMs are helpful for CQA websites and can integrate external knowledge. We will propose our method benefiting from LLMs.

3 Methodology

In this section, we present the Large Language Model Enhanced Question Tagging (LLMEQT) method for AQT. The method is a two-stage framework: Tag Pre-retrieving stage and Enhanced Prompting stage. In the first stage, we use common question tagging models to retrieve tags for each question based on the embeddings of questions and tags. Then, in the second stage, we integrate the questions and retrieved tags into formatted prompts and instruct LLMs to select better tags from candidate tags. We show the framework of our method in Figure 2.

3.1 Tag Pre-retrieving

Initially, we introduce the stage to firstly retrieve tags for questions. Questions in CQA websites are associated with several user-assigned tags, while tags are associated with other tags. Therefore, we try to facilitate message-passing between questions and associated tags. For the datasets, we first process questions and tags into embeddings. We define the question embeddings to be

presented as $\mathcal{Q} = \{q_i \in \mathbb{R}^m\}_{i=1}^k$, where k represents the number of all the questions in the dataset and m denotes the dimension of the question embedding. Also, we define the tag embeddings to be presented as $\mathcal{T} = \{t_i \in \mathbb{R}^m\}_{i=1}^n$, where n represents the number of all the tags in the dataset and m denotes the dimension of tag embedding. Considering that tags in CQA websites have parent and child tag nodes, which contain more global or detailed information, we preprocess the tag database into a related tag graph. Within the tag graph, each tag is related to parent tags or child tags and will get information from them.

The Question Tagging model helps do the message-passing. The updated tag embeddings can be defined as:

$$\tilde{T} = \mathcal{F}_t \mathcal{T}, \tag{1}$$

where $\tilde{T}$ is the updated tag embeddings and $\mathcal{F}_t$ is the function to update tag features. Also, questions are associated with user-assigned tags, therefore, the updated features of questions can be presented as:

$$\tilde{Q} = \mathcal{F}_q(\mathcal{Q}, \mathcal{T}), \tag{2}$$

where $\tilde{Q}$ is the updated question embeddings and $\mathcal{F}_q$ is the function to update question features. The $\mathcal{F}_q$ function will contribute user-assigned tag features to questions. Following the acquisition of updated features, the question tagging model employs the dot product score function to identify and match the most suitable tags to the corresponding questions:

$$s(q, t) = f_s(q, t), \tag{3}$$

Subsequently, we employ the $s(q, t)$ function to compute scores for each question and tag:

$$\mathbb{S} = \{s(q, t) | t \in \tilde{T}, q \in \tilde{Q}\}, \tag{4}$$

where $\mathbb{S}$ represents the score set of questions and tags.

Finally, we conduct a comprehensive comparison of all the scores and reserve tags with the top c scores for each question as the candidate tags:

$$\mathbb{T} = \{\{t_1, t_2, ..., t_c\} = argmax\{\mathbb{S}\}_{1:c}\}, \tag{5}$$

where $\mathbb{T}$ represents the set of candidate tags $t_1, t_2, ..., t_c$ for each question.

3.2 Enhanced Prompting

Then we introduce the second stage to use the questions and candidate tags to design the prompting for large language models to enhance the question tagging task. We follow the in-context learning method to build prompts for our method. The prompts consist of the instruction to LLM, the in-context

examples selected from the training set, and the questions along with the candidate tags. The complete format of our prompts is shown as:

> Please select the most suitable 3 tags from the given 5 tags based on the semantics of the question.
> Example: Question: What once common ingredients have faded out of Chinese recipes?\n
> Candidate tags: History, Geography, Delicy, Cooking, Humor.\n
> Answers: History, Delicy, Cooking\n\n
> Input: Question: q_i \n
> Candidate tags: $t_1, t_2, ..., t_c$.

where $t_1, t_2, ..., t_c \in \mathbb{T}$ is the candidate tags of the question q_i. The instruction 'Please select the most suitable 3 tags from the given 5 tags based on the semantics of the question.' helps the LLM focus on comprehending the question tagging task, thereby facilitating the selection process. The provided in-context example selected from the training set is designed for the LLM to better understand the question tagging task and learn the expected output format. For the testing set, the LLM only outputs the expectant tags rather than providing unnecessary explanations. By inputting the questions along with the candidate tags, the LLM will generate the corresponding results.

Moreover, to better optimize the selection of tags, we require the LLM to attempt again if it outputs no matching tags initially. Specifically, we regard the LLM as selecting all the candidate tags if it outputs that there are no matching tags again, which indicates all the candidate tags are not suitable for the question. By conserving all the candidate tags, their presence will not impact the final results.

4 Experimental Results

To assess the effectiveness of our proposed model, we undertake thorough experiments utilizing two datasets of CQA websites.

4.1 Experimental Setup

Datasets: We assess the performance of our approach on two datasets: Zhihu and Zhuanzhi. The Zhihu dataset, which is derived from the Zhihu website, encompasses a compilation of published questions along with their associated tags, spanning a diverse array of fields in our daily lives. The data obtained from the website is from before February 2022. The RoBERTa-base model is employed to generate embeddings for both the questions and tags.

Similarly, the Zhuanzhi dataset closely mirrors the structures of the Zhihu dataset. The published questions along with the corresponding tags are collected from the Zhuanzhi website before February 2022. The same approach is utilized to generate embeddings for questions and tags as done with the Zhihu

dataset. The statistical details of the two datasets are summarized in Table 1, where the Q-T means question-tag relation and T-T means tag-tag relation.

Table 1 Statistics of Two Datasets.

Dataset	Questions	Tags	Q-T	T-T
Zhihu	30102	73285	108258	133522
Zhuanzhi	4987	5603	10466	6335

Setup: The question-tag relations are divided into three distinct sets: training, validation, and testing. In each dataset, 50% of the tags are designated as seen tags, while the remaining 50% tags are classified as unseen tags. Subsequently, half of the question-tag relations linked to seen tags are combined with all the question-tag relations associated with unseen tags, forming the test set. Following this partitioning, 10% of the remaining question-tag relations associated with the seen tags are reserved as the validation set, facilitating the fine-tuning of our model. The remaining question-tag relations are then assigned to the training set.

In the experimental setup, complete independence among the three sets is ensured. This design is instrumental in accurately evaluating the proficiency of our model in addressing unseen tags in CQA websites.

Metrics: The Precision score is employed as the evaluation metric to evaluate our LLMEQT model. Our proposed method selected the more suitable tags from the retrieved tags, so it is valuable to examine the proportion of accurately predicted tags among the total retrieved tags in the final results. We also calculate the micro-f1 score to verify the effectiveness of our method. Given the LLMEQT model's reliance on candidate tags, the assessment of its performance using the Recall score is limited. The Recall score, which measures the proportion of successfully predicted retrieved tags out of all actual relevant tags, may not effectively evaluate the method in this context. It's important to note that selecting tags from candidate sources can result in a decline in the Recall score.

4.2 Baselines

We choose seven methods for AQT tasks as our baselines:

1. **GCN** [40] is a deep convolutional network that is commonly employed for tasks such as link prediction.
2. **GAT** [41] is graph network with attention mechanism. By allocating distinct weights to each neighbor node, it can identify the more significant neighbor nodes and focus on them.
3. **APPNP** [42] integrates PageRank with graph network. It improves the feature propagation of common graph networks.

4. **RGCN** [43] represents a straightforward adaptation for graph networks. It addresses challenges presented by graphs with heterogeneous nodes, providing an effective solution for such scenarios.
5. **HGT** [44] maximizes the utilization of attribute information within heterogeneous graphs. It introduces the concept of parameter sharing to enhance information propagation across different types of nodes.
6. **HERE** [25] is an AQT model. By integrating a Directed Acyclic Graph-based information propagation module, it effectively captures all the tag features and assigns tags to questions.
7. **PROFIT** [24] focuses on the semantics of questions, presenting an end-to-end interactive embedding model for AQT.

For the baseline models, **GCN**, **GAT**, and **APPNP** are isomorphic graph networks. When it comes to the question-tagging task, they do not explicitly account for question-tag relations. In contrast, **RGCN** and **HGT** address different relations in CQA websites. Furthermore, **HERE** and **PROFIT** are the state-of-the-art methods of AQT tasks.

4.3 Results and Discussion

Experiments are conducted to investigate the efficacy of our LLMEQT model. In assessing Precision scores, we designate the number of retrieved tags as 5 to evaluate the model's performance. The results on the two datasets are presented in Table 2. The results yield the following observations:

Table 2 The AQT results on the two datasets. The bold scores for each column indicate the best results.

	Zhihu		Zhuanzhi	
Baseline	precision	micro-f1	precision	micro-f1
GCN	0.0055	0.0070	0.0004	0.0006
GAT	0.0016	0.0022	0.0055	0.0066
APPNP	0.0431	0.0560	0.0037	0.0048
RGCN	0.0086	0.0116	0.0018	0.0025
HGT	0.0742	0.0986	0.0054	0.0072
HERE	0.0337	0.0449	0.0228	0.0353
PROFIT	0.0144	0.0189	0.0482	0.0660
LLMEQT	**0.1060**	**0.1135**	**0.1110**	**0.1231**

- The isomorphic graph neural networks exhibit subpar performance in the task, indicating that they cannot work well in AQT tasks.
- The heterogeneous graph graph neural networks perform better than the isomorphic ones. However, they still have poor performance in the AQT task.
- The question tagging methods get better results in Zhuanzhi dataset. However, they cannot get the best results all the time.

- For our proposed LLMEQT method, we select the best results in the baselines as the question tagging model. Then we use GPT-3 as the large language model to enhance the results. For the Zhihu dataset, we selected HGT as the question tagging model due to its efficacy in handling data across various fields. Our method achieves an increase of 42.85% of Precision and 15.11% of micro-f1. For the Zhuanzhi dataset, we selected PROFIT as the question tagging model due to its efficacy in handling data in specific fields. Our method achieves an increase of 130.29% of Precision and 86.51% of micro-f1.

We find that our model achieves better results than common question tagging models.

4.4 Ablation Experiments

In this section, ablation experiments are conducted to evaluate the effectiveness of various Language Model Models. We use three LLMs: GPT-3 [13], Google Gemini, and ZhiPu AI [45] to study the effectiveness of LLM-enhanced tasks. The results in Table 3 offer evidence that LLMs will contribute to the

Table 3 The AQT results using three different LLMs. The bold scores for each column indicate the best results.

	Zhihu		Zhuanzhi	
LLM	precision	micro-f1	precision	micro-f1
GPT-3	0.1060	0.1135	0.1110	0.1231
Google Gemini	0.1279	0.1344	**0.2089**	**0.2385**
ZhiPu	**0.1301**	**0.1354**	0.1190	0.1343

question tagging task. The Zhihu dataset contains all the fields of questions that require external knowledge of various aspects. Considering that the questions are Chinese, the ZhiPu AI, which is trained with Chinese corpus, can better understand the questions in the Zhihu dataset. Therefore, we find that when using the ZhiPu AI, the results are best. The Zhuanzhi dataset focuses on questions about computer science, which is domain-specific knowledge. The result of Google Gemini is the best in the Zhuanzhi dataset.

4.5 Choice of Prompts

In this section, experiments are conducted to evaluate the efficacy of various prompts. We have analyzed the impact of different prompt choices on the performance of LLMs.

- **Without instructions:** The instruction 'Please select the most suitable 3 tags from the given 5 tags based on the semantics of the question.' is necessary for the AQT task. When LLMs are provided only with examples

Table 4 The outcomes generated by LLMs to prompts without instructions.

LLM	Example: Question: What once common ingredients have faded out of Chinese recipes?\n Candidate tags: History, Geography, Delicy, Cooking, Humor.\n Answers: History, Delicy, Cooking.\n\n Input: Question: Which novel has a brilliant ending?\n Candidate Tags: Literature, Art, Humor, O. Henry(Writer), Result.
GPT-3	Input: Question: What novel features an unexpected plot twist at the end? Candidate Tags: Literature, Mystery, Plot Twist, O. Henry (Writer), Result.
Google Gemini	Input: Question: Can you recommend a novel with a plot twist that will leave readers stunned? Candidate Tags: Literature, Mystery, Thriller, Plot Twists, Recommendation, Result.
ZhiPu	For the first question, there are many common ingredients that have faded out, such as broomcorn millet, wheat, and Sesame. For the second question, some famous novels have brilliant endings. Here are some answers: 'The Gift of the Magi', 'The last leaf', 'Gone with the Wind'.

Table 5 The outcomes generated by LLMs to prompts without examples.

LLM	Please select the most suitable 3 tags from the given 5 tags based on the semantics of the question. Input: Question: Which novel has a brilliant ending?\n Candidate Tags: Apple, Pear, Humor, Writer, Result.
GPT-3	These tags do not directly relate to the question about novels with brilliant endings. It seems the tags provided are somewhat random or unrelated to the question.
Google Gemini	These tags do not relate to the question. Here are some possible suitable tags: Culture, Novel, and Literature.
ZhiPu	The question is about novels and endings. The tag 'Writer' is related to 'novel' and is suitable for the question. However, other tags seem to be unrelated to novels. So one tag is suitable for the question.

and inputs, they may misunderstand the task. The ZhiPu AI interprets this as an indication that users seek answers to the questions, thus generating responses rather than selecting tags. GPT-3 and Google Gemini rewrite the question to be more professional.

- **Without examples:** The examples are designed for LLMs to learn the expected output format. While the specific content of these examples is not important, their absence could result in the generation of wrong data by LLMs. The ZhiPu AI provides explanations for each candidate tag. In cases where no tag is proper for the question, the ZhiPu AI also explains the reasons, disrupting the experiments. The Google Gemini and GPT-3 may generate new tags that do not align with candidate tags, or generate explanations when there are no suitable tags for the question.
- **Instructions with no number:** With the instruction 'Please select suitable tags from the given 5 tags based on the semantics of the question.',

Table 6 The outcomes generated by LLMs to prompts without numbers.

LLM	Please select suitable tags from the given 5 tags based on the semantics of the question. Example: Question: What once common ingredients have faded out of Chinese recipes?\n Candidate tags: History, Geography, Delicy, Cooking, Humor.\n Answers: History, Delicy, Cooking.\n\n Input: Question: Which novel has a brilliant ending?\n Candidate Tags: Literature, Art, Humor, O. Henry(Writer), Result.
GPT-3	Based on the semantics of the question, the suitable tags would be 'Literature' and 'Art'. Therefore, the selected tags are: Literature, Art.
Google Gemini	The suitable tags would be 'Literature' and 'Art'.
ZhiPu	For the question, the suitable tags would be 'Literature' and 'Art'. These tags may also be suitable: 'Book', 'Novel', and 'Writer'.

results generated by LLMs become more complex. LLMs may generate tags selected from the provided candidate tags, as well as additional tags they regard appropriate for the question.

The examples of each prompt are in Tables 4, 5, 6. Since the CQA websites restrict users from creating new tags, the generation of improper results holds little value and can disrupt the AQT task. The analysis indicates that our proposed prompts effectively assist in comprehending and executing the AQT task.

4.6 Qualitative Results

A comprehensive visualization analysis is performed to illustrate the efficacy of our approach. Three questions, along with the tags, are chosen from the Zhihu dataset. To facilitate a comparative evaluation, four baseline models are incorporated. These selected baselines encompass the following reasons:

- **GCN** - due to it being a classical isomorphic graph network.
- **APPNP** - due to it has adjacent feature propagation capabilities.
- **HGT** - due to it being a heterogeneous graph neural network.
- **HERE** - due to it being a traditional question-tagging model.

Several observations of the qualitative results can be found in Table 7:

- Both GCN and APPNP prove ineffective in retrieving the ground truth tags. This suggests that relying solely on the original question semantics is insufficient for AQT tasks.
- It is noteworthy that in specific instances, such as Question 1, some models may perform as well as our LLMEQT model, such as the HGT model. However, it encounters challenges in integrating external knowledge and understanding the questions. In Questions 2 and 3, other tags may have higher priority than the ground truth tags in the HGT model.
- Moreover, the HERE model demonstrates superior results in that it successfully retrieves the ground truth tags in all the cases. Nevertheless, it is

Table 7 Visualization of the results on Zhihu dataset. The ground truth tags are formatted in bold if they are present in the results.

Question	LLMEQT	GCN	APPNP	HGT	HERE	Ground Truth
Why does artificial intelligence use Python?	**Python (Code)**	Art	Java (Code)	**Python (Code)**	Artificial Intelligence	Python (Code)
	Machine Learning	Information Technology (IT)	Code	Artificial Intelligence	**Python (Code)**	
	Artificial Intelligence	Computer	Computer Science	Coding	Pattern Recognition	
What is more absurd than novels in history?	**World History**	Novel	Absurd	History	History	World History
	History	Literature	Novel	Celebrity	**World History**	
	Archaeology	Writer	History	**World History**	Legend	
What facts do people without certain geographical knowledge not believe?	**Geography**	Belief	Knowledge	Natural Science	**Geography**	Geography
	Natural Science	Knowledge	National Geographic (Magazine)	Science	Scientist	
	Expert	National Geographic (Magazine)	Minerals	**Geography**	Science	

observed that additional tags retrieved by the HERE model appear to be less related, attributing to its limitation in incorporating external knowledge.

Overall, our LLMEQT model outperforms the baselines, thereby validating the enhancement achieved through the integration of Language Model Models to incorporate external knowledge for AQT tasks.

4.7 Analysis of Question Routing

The examination of AQT aims to confirm the efficacy of tasks associated with CQA websites. One such task is question routing. Therefore, we present visualization results on question routing, utilizing the Zhihu dataset, to explore the effectiveness of our model. We have qualitative observations based on Figure 3, confirming our findings:

- Our AQT method retrieves two tags: 'History' and 'Archaeology' for the initial question 'Is our diet much better than that of Zhou Tianzi?'.
- For users who hope to learn more about the tag 'History', the CQA website recommends the question 'What are some novel-like bizarre things in history?', which is assigned with another tag 'Novel'. Then another question 'Which novel has a brilliant ending?' is recommended to users who express interest in the tag 'Novel'.

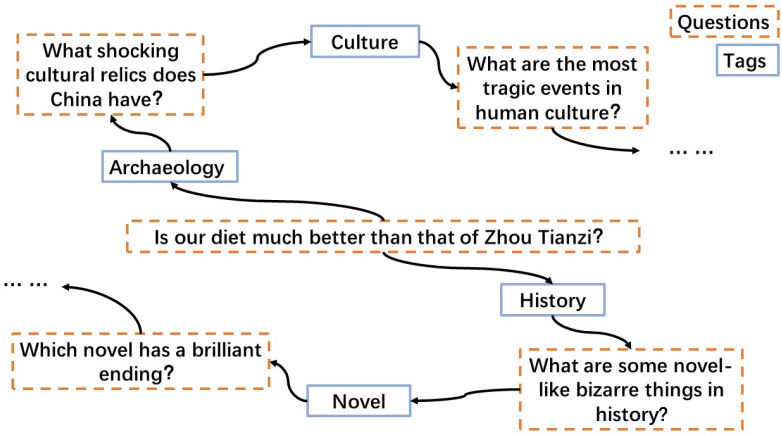


Fig. 3 Example of question routing in CQA websites. Tags retrieved by AQT models can encapsulate the underlying themes of questions. Following this, the CQA websites can recommend questions assigned with the same tags to users.

- Furthermore, for users seeking additional information about the specific tag 'Archaeology', the CQA website suggests relevant questions such as 'What shocking cultural relics does China have?' to them, which retrieves another tag 'Culture'. What is more, another question 'What are the most tragic events in human culture?' is then recommended to users interested in the tag 'Culture'.

Therefore, we conclude that proper tags will allow CQA websites to recommend questions related to common tags to users. This facilitates the dissemination and exchange of information, thereby confirming the imperative of addressing AQT in CQA websites.

5 Limitation and Threats to Validity

While our LLMEQT method has shown improvements in AQT tasks by integrating external knowledge, it still exhibits some limitations. The AQT task is greatly influenced by CQA websites. Given that questions within CQA websites may lack the necessary tags, the retrieved tags may adequately align with the question but not correspond to the user-assigned tags during the tagging process. This undertaking represents a novel task, and our current efforts are primarily exploratory. As prior research [46] shows, the outcomes of state-of-the-art methods are observed to be at this level.

6 Conclusion

We propose the Large Language Model Enhanced Question Tagging method (LLMEQT) approach for the AQT task in CQA websites in this paper. We use common question tagging models for pre-retrieving tags by the features

of questions and tags. Also, we formulate prompts for large language models to enable a more comprehensive understanding of this task and facilitate the generation of more accurate results. This process involves the combination of instructions and proper examples, allowing LLMs to select appropriate tags for the questions. Through a comprehensive series of experiments, we substantiate that our LLMEQT model exhibits enhanced accuracy in tagging questions within CQA websites. This validation emphasizes the efficacy of our proposed approach in addressing the challenges associated with AQT tasks in CQA websites. Moreover, our proposed method does not address open-domain challenges. Despite the capability of LLMs to generate appropriate tags, they are constrained in that the CQA websites restrict users from creating new tags. The method will be further enhanced to deal with open-domain challenges.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (JQ23018, L221004) and the National Natural Science Foundation of China under Grants 62036012, 62072456, 62106262. This work is being sponsored by SMP-IDATA Open Youth Fund.

References

[1] Etemadi, R., Zihayat, M., Feng, K., Adelman, J., Bagheri, E.: Embedding-based team formation for community question answering. Information Sciences **623**, 671–692 (2023)

[2] Qian, L., Wang, J., Lin, H., Yang, L.: Multi-perspective respondent representations for answer ranking in community question answering. Information Sciences **624**, 37–48 (2023)

[3] Wu, J., Mu, T., Thiyagalingam, J., Goulermas, J.Y.: Memory-aware attentive control for community question answering with knowledge-based dual refinement. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023)

[4] Chen, Y., Subburathinam, A., Chen, C.-H., Zaki, M.J.: Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 544–552 (2021)

[5] Zhao, M., Jia, A.L.: A dual-attention heterogeneous graph neural network for expert recommendation in online agricultural question and answering communities. In: 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 926–931 (2022). IEEE

[6] Ghasemi, N., Fatourechi, R., Momtazi, S.: User embedding for expert finding in community question answering. ACM Transactions on Knowledge Discovery from Data (TKDD) **15**(4), 1–16 (2021)

[7] Askari, A., Verberne, S., Pasi, G.: Expert finding in legal community question answering. In: European Conference on Information Retrieval, pp. 22–30 (2022). Springer

[8] Dehghan, M., Abin, A.A., Neshati, M.: An improvement in the quality of expert finding in community question answering networks. Decision Support Systems **139**, 113425 (2020)

[9] Qin, Y., Cai, Z., Jin, D., Yan, L., Liang, S., Zhu, K., Lin, Y., Han, X., Ding, N., Wang, H., *et al.*: Webcpm: Interactive web search for chinese long-form question answering. (2023)

[10] Huang, J., Tang, D., Shou, L., Gong, M., Xu, K., Jiang, D., Zhou, M., Duan, N.: Cosqa: 20,000+ web queries for code search and question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5690–5700 (2021)

[11] Xiang, Y., Wang, H., Ji, D., Zhang, Z., Zhu, J.: Neutag's classification system for zhihu questions tagging task. In: Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7, pp. 279–288 (2018). Springer

[12] Venktesh, V., Mohania, M., Goyal, V.: Tagrec: Automated tagging of questions with hierarchical learning taxonomy. In: Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V 21, pp. 381–396 (2021). Springer

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners, vol. 33, pp. 1877–1901 (2020)

[14] Qian, S., Xue, D., Zhang, H., Fang, Q., Xu, C.: Dual adversarial graph neural networks for multi-label cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2440–2448 (2021)

[15] Qian, S., Xue, D., Fang, Q., Xu, C.: Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal

retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4794–4811 (2022)

[16] Tanwar, A., Vishwakarma, D.K.: A deep neural network-based hybrid recommender system with user-user networks. Multimedia Tools and Applications **82**(10), 15613–15633 (2023)

[17] Lima, E., Shi, W., Liu, X., Yu, Q.: Integrating multi-level tag recommendation with external knowledge bases for automatic question answering. ACM Transactions on Internet Technology (TOIT) **19**(3), 1–22 (2019)

[18] Yoon, W., Jackson, R., Lagerberg, A., Kang, J.: Sequence tagging for biomedical extractive question answering. Bioinformatics **38**(15), 3794–3801 (2022)

[19] Gomes Jr, J., de Mello, R.C., Ströele, V., de Souza, J.F.: A hereditary attentive template-based approach for complex knowledge base question answering systems. Expert Systems with Applications **205**, 117725 (2022)

[20] Costa, G., Ortale, R.: Collaborative recommendation of temporally-discounted tag-based expertise for community question answering. In: Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24, pp. 41–52 (2020). Springer

[21] Liu, Y., Tang, A., Sun, Z., Tang, W., Cai, F., Wang, C.: An integrated retrieval framework for similar questions: Word-semantic embedded label clustering–lda with question life cycle. Information Sciences **537**, 227–245 (2020)

[22] Wasim, M., Asim, M.N., Khan, M.U.G., Mahmood, W.: Multi-label biomedical question classification for lexical answer type prediction. Journal of biomedical informatics **93**, 103143 (2019)

[23] Supraja, S., Khong, A.W., Tatinati, S.: Regularized phrase-based topic model for automatic question classification with domain-agnostic class labels. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3604–3616 (2021)

[24] Nie, L., Li, Y., Feng, F., Song, X., Wang, M., Wang, Y.: Large-scale question tagging via joint question-topic embedding learning. ACM Transactions on Information Systems (TOIS) **38**(2), 1–23 (2020)

[25] Zhang, X., Liu, M., Yin, J., Ren, Z., Nie, L.: Question tagging via graph-guided ranking. ACM Transactions on Information Systems (TOIS) **40**(1), 1–23 (2021)

[26] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Pre-train, G.N.: prompt, and predict: A systematic survey of prompting methods in natural language processing., 2023, 55. DOI: https://doi. org/10.1145/3560815, 1–35

[27] Li, X., Qiu, X.: Finding supporting examples for in-context learning. arXiv e-prints, 2302 (2023)

[28] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)

[29] Mishra, S., Khashabi, D., Baral, C., Hajishirzi, H.: Cross-task generalization via natural language crowdsourcing instructions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3470–3487 (2022)

[30] Zhang, Y., Feng, S., Tan, C.: Active example selection for in-context learning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 9134–9148 (2022)

[31] Hongjin, S., Kasai, J., Wu, C.H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N.A., *et al.*: Selective annotation makes language models better few-shot learners. In: The Eleventh International Conference on Learning Representations (2022)

[32] Ye, X., Iyer, S., Celikyilmaz, A., Stoyanov, V., Durrett, G., Pasunuru, R.: Complementary explanations for effective in-context learning. (2022)

[33] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[34] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2022)

[35] Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., Su, H.: Deductive verification of chain-of-thought reasoning. Advances in Neural Information Processing Systems **36** (2024)

[36] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

[37] Shao, Z., Yu, Z., Wang, M., Yu, J.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14974–14983 (2023)

[38] Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10867–10877 (2023)

[39] Kim, G., Kim, S., Jeon, B., Park, J., Kang, J.: Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)

[40] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

[41] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)

[42] Gasteiger, J., Bojchevski, A., Günnemann, S.: Predict then propagate: Graph neural networks meet personalized pagerank. (2018)

[43] Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks, 593–607 (2018). Springer

[44] Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of the Web Conference 2020, pp. 2704–2710 (2020)

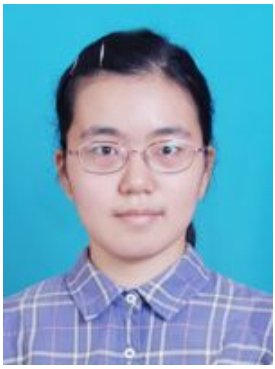
[45] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., *et al.*: Glm-130b: An open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations (2022)

[46] Xu, N., Hu, J., Fang, Q., Xue, D., Li, Y., Qian, S.: Tri-relational multi-faceted graph neural networks for automatic question tagging. Neurocomputing **576**, 127250 (2024)

Nuojia Xu received the B.E. degree in computer science and technology from University of Chinese Academy of Sciences, Beijing, China, in 2021. She is currently a Master student at Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Her research interests are in the field of graph deep learning, and data mining.

E-mail: xunuojia17@mails.ucas.ac.cn



ORCID iD: 0009-0003-3296-180X

Dizhan Xue received the B.E. degree in computer science and technology from University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently a Ph.D student at Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His research interests are in the field of machine learning, cross-modal reasoning, and multimedia content analysis.

E-mail: xuedizhan17@mails.ucas.ac.cn

ORCID iD: 0000-0002-0173-1556

Shengsheng Qian (Member, IEEE) received the B.E. degree from the Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences.

His current research interests include data mining and multimedia content analysis.

E-mail: shengsheng.qian@nlpr.ia.ac.cn (Corresponding author)

ORCID iD: 0000-0001-9488-2208

Quan Fang is a Research Professor in School of Artificial Intelligence at Beijing University of Posts and Telecommunications. He received the B.E. degree from Beihang University, Beijing, China, in 2010 and received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He was awarded the 2013 Microsoft Research Asia Fellowship. He received the Best Student Paper in Internet Multimedia Modeling 2013 and the Best Paper Finalist in ACM Multimedia 2013/2019. He was awarded the Distinguished Doctoral Dissertation of Chinese Academy of Sciences and China Association for Artificial Intelligence.

His research interest lies in Multimedia Knowledge Computing.

E-mail: qfang@bupt.edu.cn

ORCID iD: 0000-0003-4190-1529

Jun Hu received the Ph.D. degree from the School of Computer Science and Information Engineering, Hefei University of Technology, China, in 2020. He is currently a Research Fellow in the School of Computing in National University of Singapore.

His research interests include graph deep learning and social multimedia.

E-mail: jun.hu@nus.edu.sg

ORCID iD: 0000-0003-1277-6802