BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 19/25 pp843–857
DOI: 10.26599/BDMA.2024.9020026
Volume 7, Number 3, September 2024

Prompting Large Language Models with Knowledge-Injection for Knowledge-Based Visual Question Answering

Zhongjian Hu, Peng Yang*, Fengyuan Liu, Yuan Meng, and Xingyu Liu

Abstract: Previous works employ the Large Language Model (LLM) like GPT-3 for knowledge-based Visual Question Answering (VQA). We argue that the inferential capacity of LLM can be enhanced through knowledge injection. Although methods that utilize knowledge graphs to enhance LLM have been explored in various tasks, they may have some limitations, such as the possibility of not being able to retrieve the required knowledge. In this paper, we introduce a novel framework for knowledge-based VQA titled "Prompting Large Language Models with Knowledge-Injection" (PLLMKI). We use vanilla VQA model to inspire the LLM and further enhance the LLM with knowledge injection. Unlike earlier approaches, we adopt the LLM for knowledge enhancement instead of relying on knowledge graphs. Furthermore, we leverage open LLMs, incurring no additional costs. In comparison to existing baselines, our approach exhibits the accuracy improvement of over 1.3 and 1.7 on two knowledge-based VQA datasets, namely OK-VQA and A-OKVQA, respectively.

Key words: visual question answering; knowledge-based visual question answering; large language model; knowledge injection

1 Introduction

Knowledge-based Visual Question Answering (VQA)[1] extends the VQA task[2], introducing the requirement for external knowledge to answer questions. Early knowledge-based VQA benchmarks also provide knowledge bases[3]. More recently, VQA benchmarks have been established that emphasizes open-domain knowledge[4, 5]. In open-domain knowledge-based VQA, any external knowledge can be applied to answer questions. This paper focuses on VQA with open-domain knowledge.

Early researchers try to retrieve knowledge from external knowledge resources. More recently, some works attempt to explore the utilization of implicit knowledge in pre-trained language models, such as KRISP[6]. With the emergence of Large Language Models (LLMs), researchers have shifted towards employing them as knowledge acquisition engines. PICa[7] adopts GPT-3[8] for in-context learning in knowledge-based VQA. Given that GPT-3, as a language model, lacks direct comprehension of images, the approach involves conversion of the image into its corresponding textual caption through a captioning model.

The VQA triplet image-question-answer will be converted to context-question-answer, thus unifying the input into text. The context denotes the caption for the image. Despite the encouraging results of PICa, we argue that the capacity of LLM can be further

• Zhongjian Hu, Peng Yang, and Yuan Meng are with the School of Computer Science and Engineering, Southeast University, and also with the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education of the People's Republic of China, Nanjing 211189, China. E-mail: huzj@seu.edu.cn; pengyang@seu.edu.cn; yuan_meng@seu.edu.cn.
• Fengyuan Liu and Xingyu Liu are with the Southeast University - Monash University Joint Graduate School (Suzhou), Southeast University, Suzhou 215125, China. E-mail: liufengyuan@seu.edu.cn; liuxingyu_025@seu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2024-02-24; revised: 2024-04-01; accepted: 2024-04-07

improved through knowledge enhancement.

Some studies have proposed to use knowledge graphs to enhance LLMs, such as RAG[9]. RAG retrieves documents from the knowledge graphs and incorporates them as supplementary contextual information. While knowledge graphs have the potential to enhance LLMs, they still have limitations, such as the possibility of not being able to retrieve the required knowledge.

Inspired by previous works, we propose a novel framework for knowledge-based VQA that enhances LLMs through knowledge injection. Figure 1 shows the comparison of PICa, RAG, and our method. In addition, we adopt the open LLMs, which are free compared to GPT-3. Existing approaches, when applying the LLM to knowledge-based VQA, only utilise the knowledge of the LLM itself, ignoring the role of external knowledge to inspire the LLM. We propose a novel framework that injects knowledge into the prompt to further inspire the LLM. In addition, to address the issue that knowledge resources such as knowledge graphs may not be able to retrieve the required knowledge, we adopt a new idea of employing another LLM to generate the knowledge instead of retrieving the knowledge from the knowledge graphs. Main contributions are as follows:

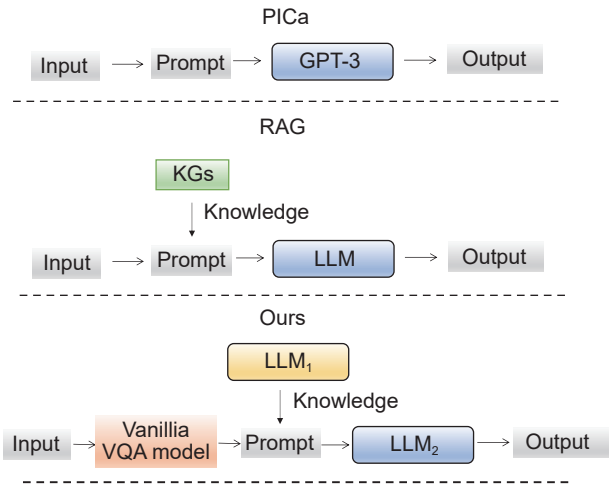


Fig. 1 Comparison of PICa, RAG, and our method. PICa applies the in-context learning of GPT-3 to knowledge-based VQA. RAG uses knowledge retrieved from the knowledge graphs to enhance the LLM. Our approach first adopts a vanilla VQA model to generate in-context examples, then takes the LLM_1 to generate background knowledge instead of retrieving knowledge from the knowledge graphs, and finally integrates the knowledge into prompt to inspire LLM_2.

- We propose a novel framework for knowledge-based VQA that incorporates appropriate in-context examples and background knowledge to predict the answer. The framework is entirely built upon open LLMs and is free of cost.
- To our knowledge, this is the first attempt to utilize LLM to enhance knowledge for another LLM in knowledge-based VQA task.
- We conduct experiments on two knowledge-based VQA datasets, namely OK-VQA and A-OKVQA. Experiments show that our approach outperforms the existing baselines.

2 Related Work

The research of Artificial Intelligence (AI) is of great significance, which not only promotes the progress of science and technology, but also has a far-reaching impact on the social and economic development, the improvement of human life and the construction of the future world. More and more AI research is emerging, which has an important impact on promoting the development of AI. SpectralGPT[10] is a remote sensing foundation model designed for spectral data. It has significant potential for advancing spectral remote sensing big data applications in geoscience across four downstream tasks: single/multi-label scene classification, semantic segmentation, and change detection. Hong et al.[11] created the C2Seg dataset for multimodal remote sensing. The C2Seg dataset is intended for use in the cross-city semantic segmentation task. To improve the generalization ability of AI models in multi-city environments, they also proposed a high-resolution domain adaptation network, referred to as HighDAN. Hong et al.[12] proposed the augmented LMM, a novel spectral mixture model, to address spectral variability in inverse problems of hyperspectral unmixing. During the imaging process, data are often affected by various variabilities. Our proposed method may also experience some limitations when faced with various variabilities. For instance, if the image is damaged, information loss may occur during the conversion of captioning. This loss of captioning information can affect knowledge generation, as it is partly based on the captioning information.

VQA. VQA[13, 14] is a popular multi-modal AI task[15, 16]. Recent VQA studies can be generally divided into these categories: better visual features[17, 18],

better model architectures[19–21], and better learning paradigms[22–24]. According to the research methods can be divided into: joint embedding methods, attention methods, modular methods, external knowledge-based methods, and so on. Most of the joint embedding methods use Convolutional Neural Network (CNN) network to extract visual features, Recurrent Neural Network (RNN) network to extract text features, and simply combine the two features. The Neural-Image-QA model proposed by Malinowski et al.[25] is the first to leverage the joint embedding method. The model is based on CNN and Long Short-Term Memory (LSTM), treating the VQA task as a sequence-to-sequence task assisted by image information. Nevertheless, the majority of joint embedding methods commonly utilize all the features extracted from both images and questions as the input for the VQA model. This approach may introduce a considerable amount of noise, potentially affecting the performance. The objective of the attention method is to concentrate the limited attention on crucial elements, significantly enhancing the comprehension capability of neural network. Yu et al.[26] introduced a multi-modal factorized bilinear pooling approach, where text attention is inferred based on the question, and visual attention is inferred by the involvement of text attention. However, the VQA task is compositional. For example, in a question like "What's on the table?", it is necessary to first determine the position of the table, then identify the location above the table, and finally ascertain the target object above the table. Hence, some studies have proposed the modular networks for VQA task. The modular approach involves designing distinct modules for various functions and connecting these modules based on different questions. Andreas et al.[27] first applied neural modular networks to VQA. Additionally, there exists a category of VQA that necessitates external knowledge, often referred to as knowledge-based VQA.

Knowledge-based VQA. Some benchmarks for knowledge-based VQA have been proposed, necessitating external knowledge to answer questions. Early works retrieve knowledge from external knowledge resources. More recently, Marino et al.[6] proposed KRISP to retrieve implicit knowledge stored in pre-trained language models. MAVEx[28] proposes a validation method aimed at improving the utilization of noisy knowledge. Yang et al.[7] proposed PICa, which applies in-context learning of GPT-3 to knowledge-based VQA, achieving encouraging results. PICa utilizes a captioning model to convert the image into corresponding caption, which can be processed by LLM. In-context learning, a powerful few-shot learning technique, enables reasoning with a few task examples assembled as the prompt, eliminating the need for parameter updates. Prophet[29] adopts a vanilla VQA model to inspire LLM, further activating the capability of LLM.

Knowledge-enhanced LLMs. LLMs have demonstrated promising results across various tasks. Researchers explore the use of knowledge graphs to enhance LLMs[30]. Knowledge graphs[31] offer a means to enhance LLMs by incorporating knowledge during pre-training, a process that extends to the inference stage as well.

When integrating knowledge graphs into training objectives, ERNIE[32] adopts a method where both sentences and corresponding entities are input into LLMs. The training process involves instructing the LLMs to predict alignment links. On the other hand, ERNIE 3.0[33] represents a knowledge graph triple as tokens, concatenating them with sentences. RAG[9] employs a distinctive approach by initially searching and retrieving relevant documents from knowledge graphs. These documents are then provided to the language model as additional context information. Despite the benefits of knowledge graphs in enhancing LLMs, they may face challenges in retrieving the required knowledge. In this paper, a novel idea is proposed, suggesting the utilization of one LLM to enhance knowledge for another LLM, as an alternative to traditional knowledge graphs.

3 Methodology

The Prompting Large Language Models with Knowledge-Injection (PLLMKI) framework is illustrated in Fig. 2. The framework comprises three main components: (1) Utilizing a vanilla VQA model to obtain in-context examples, which are then processed by a captioning model to transform the image-question-answer into context-question-answer. (2) Employing the LLM_1 to generate background knowledge and integrating the knowledge into the prompt, resulting in context-question-knowledge-answer. (3) Inputting the modified prompt into the LLM_2 to predict the answer.

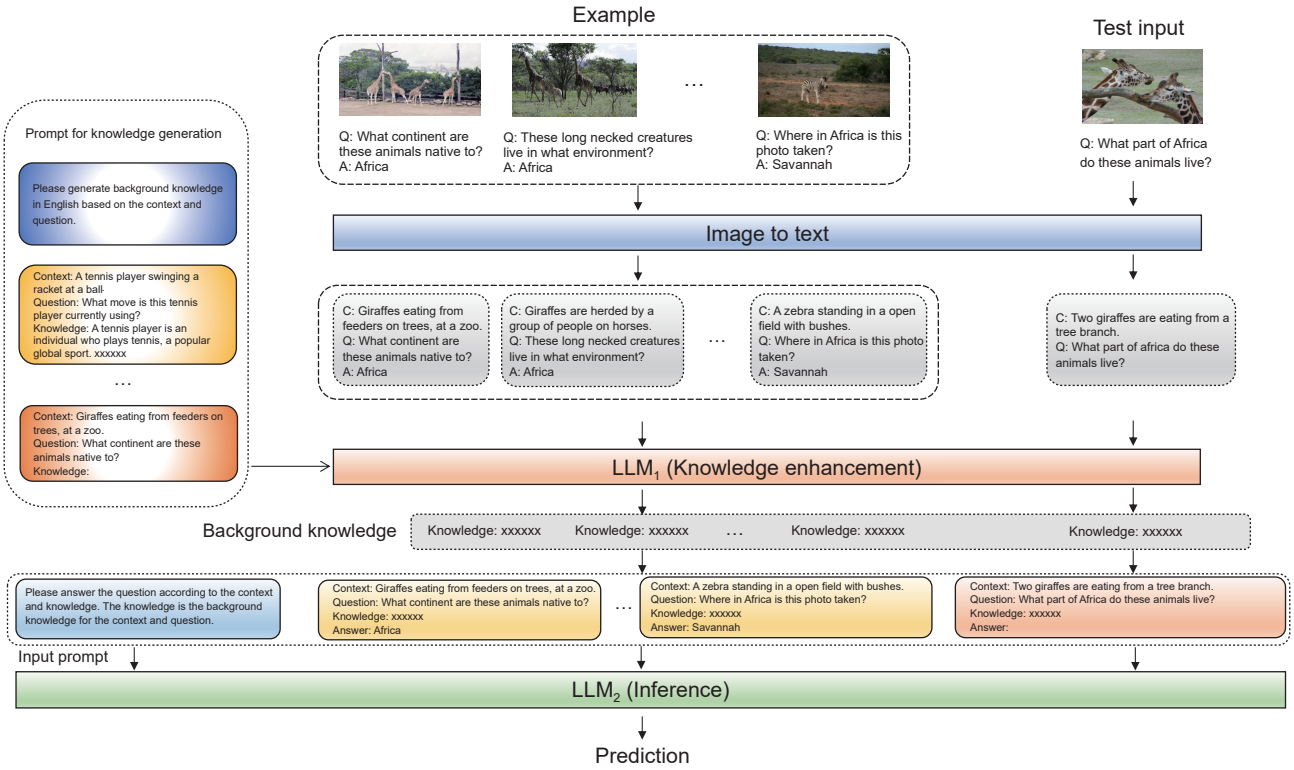


Fig. 2 Overview of the proposed framework.

3.1 Preliminary

Before introducing the PLLMKI framework, let us first demonstrate PICa. PICa leverages the in-context learning of GPT-3 for knowledge-based VQA.

The in-context learning paradigm of GPT-3 demonstrates the capable learning ability. Specifically, target y is predicted conditioned on prompt (σ, ϵ, x), where σ is the prompt head, ϵ denotes the in-context examples, and x refers to test input. At each decoding step l:

$$y^l = \arg\max_{y^l} p_{\text{LLM}}\left(y^l \mid \sigma, \epsilon, x, y^{<l}\right) \tag{1}$$

where the prompt head σ is the task description, $\epsilon = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ denotes the n in-context examples. The (x_i, y_i) denotes the input-target pair for the task.

PICa applies the in-context learning paradigm of GPT-3 to VQA. Since GPT-3 lacks the ability to comprehend image input, the image is first converted to a caption using the image captioning model. The original triplet from the VQA dataset, consisting of image-question-answer, is transformed into context-question-answer format. PICa defines the in-context example in the format:

Context: the caption.

Question: the question.

Answer: the answer.

The context-question-answer triplet corresponds to the image-question-answer from the training set. The test input is structured in the format:

Context: the caption.

Question: the question.

Answer:.

The format of the test input mirrors that of the in-context example, with the exception that the answer is left blank, allowing the LLM to make predictions. PICa defines the prompt head as follows: Please answer the question according to the above context.

The prompt head, in-context examples, and test input are assembled together to form the whole prompt. The prompt is then fed into the LLM, such as GPT-3, to predict the answer.

The PLLMKI framework draws inspiration from previous works. We enhance the LLM through knowledge injection. Specifically, we generate background knowledge by prompting the LLM_1, and then integrate the knowledge into the prompt to assist the LLM_2 in making predictions. The in-context example is formatted as follows:

Context: the caption.
Question: the question.
Knowledge: the background knowledge.
Answer: the answer.
We format the test input as follows:
Context: the caption.
Question: the question.
Knowledge: the background knowledge.
Answer:.

The test input follows a format similar to the in-context example, with the exception that the answer is left blank for the LLM to predict. Additionally, we modify the prompt head to enable the model to answer questions based on the context and knowledge. Notably, our approach differs from previous works as the knowledge is generated by prompting another LLM, rather than being derived from knowledge graphs.

3.2 In-context example selection

Existing works[7, 34] have shown the importance of in-context example selection. We denote the image-question-answer triplet as (v,q,a), where v denotes the image, q denotes the question, and a denotes the answer. The VQA dataset is denoted as D, and the model learned from D is denoted as M.

We can obtain the fused feature $\mathcal{F}$ of the image and the question through the encoder of the model M:

$$\mathcal{F} = M(v,q) \tag{2}$$

For each test input, the cosine similarity of the fused feature between the test input and each training sample is calculated. We then select the samples from the training set whose fused features are closest to that of the test input to form the in-context examples ϵ:

$$\epsilon = \{(v_i,q_i,a_i)|_{i=1}^{n}\} \tag{3}$$

The image is converted to the context by the captioning model, and the image-question-answer corresponds to context-question-answer. Denote context-question-answer as $(c,\ q,\ a)$, where c, q, and a refer to context, question, and answer, respectively.

$$\epsilon = \{(c_i,q_i,a_i)|_{i=1}^{n}\} \tag{4}$$

3.3 Knowledge enhancement

Previous works[30] have demonstrated the benefits of enhancing LLMs with knowledge graphs. Unlike previous works, we use another LLM to inject knowledge instead of knowledge graphs. This idea is motivated by the belief that implicit knowledge embedded in LLMs may be more suitable for open-domain knowledge-based VQA compared to knowledge graphs, which may fail to retrieve the required knowledge.

We generate background knowledge by prompting LLM_1 and subsequently integrate the knowledge into the prompt to aid the LLM_2 in making predictions. The format of the prompt to generate background knowledge is as follows: {[# Prompt head] Please generate the background knowledge xxxxxx. [# In-context examples] Context: the caption. Question: the question. Knowledge: the background knowledge. [# Input] Context: the caption. Question: the question. Knowledge: }.

The prompt includes the prompt head $\mathcal{H}$, the in-context examples $\mathcal{E}$, and the test input $\mathcal{T}$. The prompt head is a task description that allows the LLM to generate background knowledge for the corresponding context and question. The in-context example is composed of context-question-knowledge and can be denoted as $(c,\ q,\ k)$.

$$\mathcal{E} = \{(c_i,q_i,k_i)|_{i=1}^{n}\} \tag{5}$$

The format of the test input resembles that of the in-context example, with the knowledge left blank for the LLM to generate.

$$\mathcal{P} = \text{LLM}_1(\mathcal{H},\mathcal{E},\mathcal{T}) \tag{6}$$

where $\mathcal{P}$ denotes the knowledge generated by the LLM.

To explain how LLM_1 can be used to generate background knowledge, we show a prompt example in Fig. 3. We pick out some in-context examples for LLM_1 learning. For better visualization, we only show one in-context example.

3.4 Prompting large language model

At this stage, we input the prompt into the LLM to predict the answer. The background knowledge generated in the previous step is integrated into the prompt, which comprises the prompt head, in-context examples, and test input: {[# Prompt head] Please answer the question xxxxxx. [# In-context examples] Context: c_i. Question: q_i. Knowledge: k_i. Answer: a_i. [# Test input] Context: c. Question: q. Knowledge: k. Answer: }

Prompt head $\mathcal{H}$ is the task description. Unlike PICa, we set the prompt head to allow the model to make

Please generate background knowledge in English based on the key words in the context and question.
===
Context: A tennis player swinging a racket at a ball.
Question: What move is this tennis player currently using?
Knowledge: A tennis player is an individual who plays tennis, a popular global sport. They use a racket to hit a ball across a net in an attempt to outmaneuver their opponent. In tennis, swinging refers to the action of moving the racket to hit the ball. The way a player swings the racket can greatly affect the trajectory, speed, and spin of the ball. In tennis, the racket is the tool that players use to hit the ball. It consists of a handle and a circular frame with tightly interwoven strings. Rackets come in various sizes and materials to fit the individual player's style and level of play. In tennis, the ball is a hollow, spherical object that players hit back and forth across a net. It is designed to have specific bounce characteristics and is covered in a fibrous felt to alter its aerodynamic properties. In the context of tennis, a move typically refers to the type of shot or stroke a player uses. There are several different moves or strokes a player can use, such as a forehand, backhand, serve, volley, overhead smash, drop shot, or lob. Each of these moves has different tactical uses in a match and requires different body positions, racket angles, and swing paths.
===
Context: A woman in a coat and boots stops to check her smartphone.
Question: What brand of purse might she be carrying?
Knowledge:

Fig. 3 Prompt example. We show the prompt with one in-context example to explain how to prompt LLM_1 to generate background knowledge. We collate several in-context examples for model learning to inspire LLM_1 to generate background knowledge.

predictions based on context and background knowledge, not just context. The in-context example $\mathcal{E}$ is also different from PICa. PICa is a triplet of context-question-answer, while ours is a quadruple of context-question-knowledge-answer.

$$\mathcal{E} = \{(c_i, q_i, k_i, a_i)|_{i=1}^{n}\} \tag{7}$$

The format of test input $\mathcal{T}$ is similar to that of the in-context example, but the answer is left blank for the LLM to predict. The prompt $(\mathcal{H}, \mathcal{E}, \mathcal{T})$ will be input into the LLM.

$$\mathcal{P} = \text{LLM}_2(\mathcal{H}, \mathcal{E}, \mathcal{T}) \tag{8}$$

where $\mathcal{P}$ is the prediction made by the LLM.

4 Experiment

4.1 Dataset

We adopt two knowledge-based VQA datasets for evaluation, namely OK-VQA[4] and A-OKVQA[5]. OK-VQA contains approximately 14 000 images and 14 000 questions, while A-OKVQA contains 24 000 images and 25 000 questions. OK-VQA encourages answering questions based on external knowledge, covering various knowledge categories such as sports, history, science, and more. A-OKVQA is also a knowledge-based VQA dataset that requires more knowledge categories than OK-VQA. For A-OKVQA, we adopt the direct answer on validation set for evaluation.

For evaluation metrics, we employ common VQA metrics. Each question is associated with ten annotated answers, and a generated answer is considered 100% accurate if at least three human annotators provided that correct answer. The accuracy metric is defined as $\min\left(\frac{\text{Number of humans that provided that answer}}{3}, 1\right)$.

4.2 Baseline and implementation

4.2.1 Baseline

We compare our approach to the following baselines:

• **MUTAN**[35] is a multimodal fusion realized by bilinear interaction. It is for modeling interactions between image and text.

• **Mucko**[36] focuses on multi-layer cross-modal knowledge inference. It represents the image as a multimodal heterogeneous graph.

• **ConceptBert**[37] learns and captures image-question-knowledge interactions from visual, language, and knowledge graph embeddings.

• **KRISP**[6] employs a multimodal Bidirectional Encoder Representations from Transformers (BERT) to process both the image and question, leveraging the implicit knowledge in BERT.

• **MAVEx**[28] uses external knowledge for multimodal answer validation. It validates the promising answers according to answer-specific knowledge retrieval.

• **Visual-retriever-reader**[38] is designed for knowledge-based VQA. The visual retriever initially fetches relevant knowledge, and subsequently, the visual reader predicts the answer based on the provided knowledge.

• **TRiG**[39] is a transform-retrieve-generate

framework that can be used with image-to-text models and knowledge bases.

• **UnifER**[40] is a knowledge-based VQA framework based on the unified end-to-end retriever-reader.

• **PICa**[7] applies the in-context learning paradigm of GPT-3 to knowledge-based VQA. It utilizes a captioning model to convert the image into text.

• **Pythia**[41] is a bottom-up top-down framework. It is improved through modifications to the model structure and the data augmentation.

• **ViLBERT**[42] is a model for learning the joint representation of image and text. It extends the BERT architecture to support multi-modality.

• **ClipCap**[43] is a captioning method that employs pre-trained models for processing visual and text.

• **LXMERT**[44] constructs a large transformer model comprising three encoders: object relationship encoder, language encoder, and cross-modality encoder.

• **GPV-2**[45] is based on General Purpose Vision (GPV) and is designed to address a wide range of visual tasks without necessitating changes to the architecture.

• **VLC-BERT**[46] is a model designed to integrate common sense knowledge into the visual language BERT.

• **Prophet**[29] is proposed to inspire GPT-3 using a vanilla VQA model. We replace GPT-3 with LLaMA and keep the settings consistent for a fair comparison.

4.2.2 Implementation

For the captioning model, we follow PICa[7], which uses OSCAR+[18]. For the in-context example selection, we follow the previous works[29] and use the MCAN-large[21] model pre-trained on VQAv2[47] and visual genome[48]. We use LLaMA1[49] as the LLM for knowledge enhancement, because LLaMA1 is an excellent and open LLM with powerful capability. We use LLaMA2[50] as the LLM for inference. Compared to LLaMA1, LLaMA2 can support a longer context, which is conducive to injecting knowledge into the context. We use the LLaMA 7B version. Considering the length limit of the context, we set the number of in-context examples to 8 and set the length of the knowledge to no more than 256. We use the default settings unless otherwise specified.

4.3 Experimental result

We report the results on OK-VQA and A-OKVQA. Tables 1 and 2 show the results.

Table 1 Results on OK-VQA.

Method	Accuracy (%)
MUTAN+AN (Ben-Younes et al.[35])	27.8
Mucko (Zhu et al.[36])	29.2
ConceptBert (Gardères et al.[37])	33.7
KRISP (Marino et al.[6])	38.9
MAVEx (Wu et al.[28])	39.4
Visual-retriever-reader (Luo et al.[38])	39.2
VLC-BERT (Ravi et al.[46])	43.1
TRiG (Gao et al.[39])	49.4
UnifER (Guo et al.[40])	42.1
PICa-Base (Yang et al.[7]) (Caption)	42.0
PICa-Base (Yang et al.[7]) (Caption+Tags)	43.3
PICa-Full (Yang et al.[7]) (Caption)	46.9
PICa-Full (Yang et al.[7]) (Caption+Tags)	48.0
Prophet-LLaMA (Shao et al.[29])	52.8
Ours	**54.1**

Table 2 Results on A-OKVQA.

Method	Accuracy (%)
Pythia (Jiang et al.[41])	25.2
ClipCap (Mokady et al.[43])	30.9
ViLBERT (Lu et al.[42])	30.6
LXMERT (Tan and Bansal[44])	30.7
KRISP (Marino et al.[6])	33.7
GPV-2 (Kamath et al.[45])	48.6
Prophet-LLaMA (Shao et al.[29])	51.2
Ours	**52.9**

On OK-VQA, our method outperforms other baselines by more than 1.3. It is evident that baselines utilizing LLMs consistently achieve better results compared to those without LLMs. LLMs are trained on extensive corpora, acquiring rich knowledge. Therefore, baselines using LLMs tend to outperform methods that do not leverage LLMs. Our approach, also grounded in LLMs, leverages the more powerful inference ability of such models, thereby outperforming the baselines that lack LLMs. Furthermore, our approach achieves better performance compared to existing LLM-based baselines. By selecting more appropriate in-context examples and further enhancing the LLM through knowledge injection, our framework demonstrates its capability to achieve superior results. On A-OKVQA, our approach consistently outperforms existing baselines, exhibiting an accuracy improvement of more than 1.7 compared to the existing baselines. These results once again underscore the effectiveness of our method.

Notably, our framework relies entirely on open LLMs, making it a cost-effective solution that is accessible to most researchers. In contrast, utilizing GPT-3 can be expensive and may result in significant costs. Our inference model employs the 7B version of LLaMA, featuring just 7 billion parameters. Notably, we achieve superior performance using only a single V100 32 G GPU for inference, a setup that is affordable for most researchers.

4.4 Ablation study

We report the results of ablation experiments conducted on both OK-VQA and A-OKVQA. The ablation experiments focus on two key aspects: (1) Knowledge enhancement under different shots: We investigate the impact of knowledge enhancement under different shots, exploring how our approach performs with different numbers of in-context examples. (2) Using different LLMs for knowledge injection: We examine the influence of employing different LLMs for knowledge injection, evaluating the performance variations.

4.4.1 Knowledge enhancement under different shots

Table 3 shows the results of ablation experiments. ✗ Knowledge denotes that no knowledge is incorporated to enhance the LLM_2. Ours (+ LLM_1) means using LLM_1 for knowledge enhancement. Ours (+ KGs) means using Knowledge Graphs (kGs)[51] for knowledge enhancement. We perform ablation experiments with and without in-context examples. 0-shot indicates no in-context examples.

Based on the results obtained from both datasets, it is evident that performance will decline in the absence of knowledge enhancement. We find that knowledge enhancement using LLM_1 outperforms KGs both with and without in-context examples, confirming our view that the implicit knowledge of LLM is more suitable for open-domain questions. In addition, in the absence of in-context examples, model performance will be significantly degraded, which also shows that in-context examples are crucial to the in-context learning paradigm of LLM.

Table 3 Ablation study.

Shot	Method	Accuracy (%)		
		OK-VQA	A-OKVQA	Average
8-shot	Ours (+ LLM_1)	**54.1**	**52.9**	**53.5**
	Ours (+ KGs)	53.6	52.1	52.9
	✗ Knowledge	52.9	51.5	52.2
0-shot	Ours (+ LLM_1)	**28.4**	**24.8**	**26.6**
	Ours (+ KGs)	27.9	24.5	26.2
	✗ Knowledge	27.2	24.2	25.7

4.4.2 Different LLMs for knowledge enhancement

Figure 4 shows the results obtained by employing various LLMs for knowledge enhancement. Here, we define LLM_1 as the LLM responsible for generating background knowledge, and LLM_2 as the LLM used for inference. The approach involves using LLM_1 to generate background knowledge, which is then injected into the prompt to guide LLM_2 during inference. Our aim is to investigate the impact of utilizing different LLMs as LLM_1. Specifically, we employ LLaMA1-7B, LLaMA1-13B, ChatGLM1, and ChatGLM2 as LLM_1, respectively, and present the corresponding experimental results.

Upon analyzing the results, it becomes apparent that more capable models tend to yield superior results. Specifically, LLaMA1-13B stands out as the top performer, attributed to its status as the largest model. The extensive parameterization of LLaMA1-13B

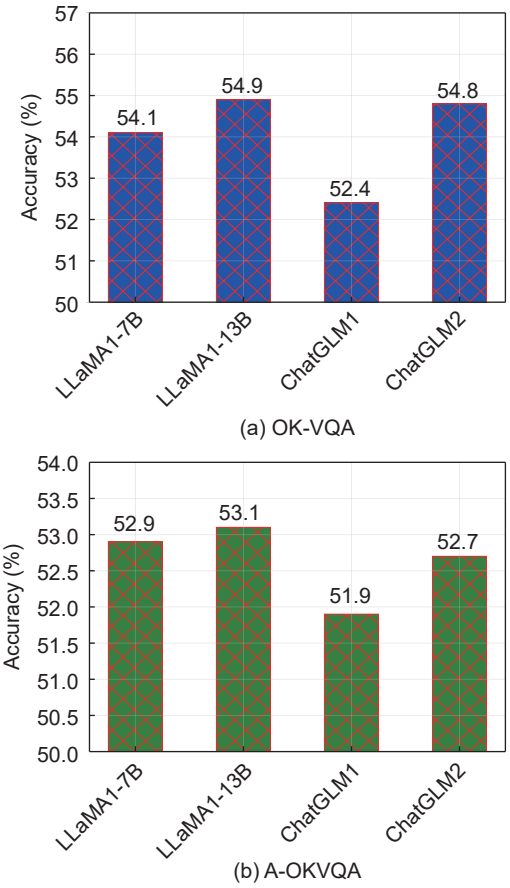


Fig. 4 Different LLMs for knowledge enhancement.

results in a richer hidden knowledge compared to other models, contributing to its superior outcomes. ChatGLM1, being trained in both Chinese and English, exhibits relatively lower performance due to its English knowledge not matching up to that of LLaMA. On the other hand, ChatGLM2, as the second-generation version of ChatGLM1, outperforms ChatGLM1, showcasing the positive impact of model advancements.

4.5 Parameter sensitivity study

We also investigate the impact of the number of in-context examples on the model performance. We analyze how performance varies with the number of in-context examples on both OK-VQA and A-OKVQA datasets. Kg-enhanced denotes the utilization of knowledge enhancement, while No Kg-enhanced signifies the absence of knowledge enhancement.

Figure 5 illustrates the variation in model performance with the number of in-context examples. Generally, the model performance exhibits an upward trend as the number of in-context examples increases. Specifically, in a 0-shot scenario (when the number of in-context examples is 0), the model performance is poor without in-context examples. However, a noticeable improvement is observed when the number of in-context examples is increased to 1, indicating the significance of in-context examples in model learning. As the number of in-context examples continues to increase, the model performance gradually reaches a saturation point, suggesting that constantly adding in-context examples does not always lead to a proportional improvement in model performance.

4.6 Case study

We select specific cases to examine the influence of background knowledge and in-context examples on model inference. The findings indicate that both background knowledge and in-context examples contribute positively to model inference.

Figure 6 illustrates the impact of background knowledge on inference. The correct answer contained in the background knowledge is denoted in blue font, while the correct answer itself is highlighted in green font. It is evident that the correct answer is present in the background knowledge, indicating that the model is more likely to make accurate predictions when leveraging background knowledge. We also present a failure case. In the last case, we can see that even if the

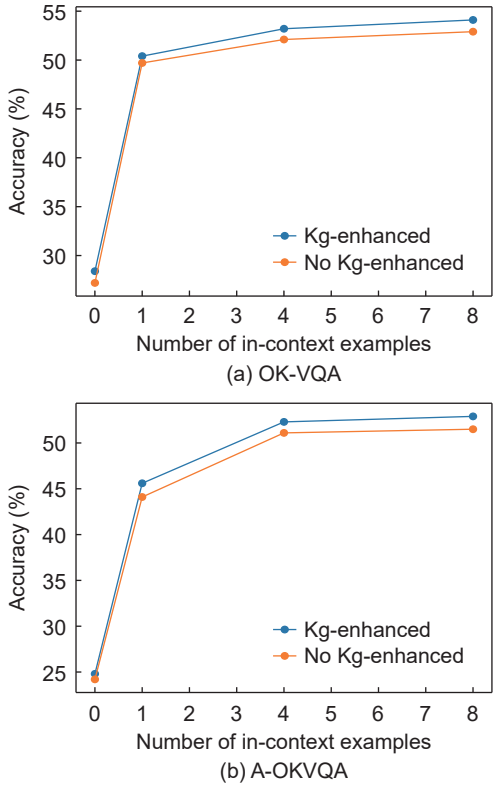


Fig. 5 Performance when varying the number of in-context examples.

correct answer is hit in the background knowledge, the model can still predict failure. This also shows that background knowledge does not always help models make 100% correct predictions.

Figure 7 depicts the impact of in-context examples on inference. The correct answer hit in the in-context example is highlighted in red font, while the correct answer itself is denoted in green font. It is evident that the correct answer is present in the in-context examples, underscoring the increased likelihood of predicting the correct answer when leveraging in-context examples.

To explain the construction details of our prompt, we show a concrete example in Fig. 8. For better visualization, we show the prompt with four in-context examples.

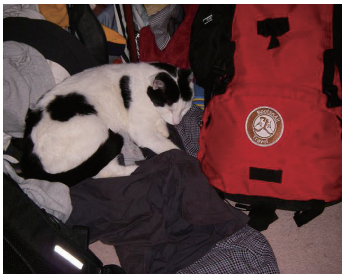
4.7 Inference time

We record the inference time for different context lengths on OK-VQA. We adopt the LLaMA2-7B version as inference model and use a Tesla V100 GPU.

The statistics for inference time are presented in Fig. 9. The *X*-axis represents time in minutes, while the *Y*-axis corresponds to different shots, indicating the

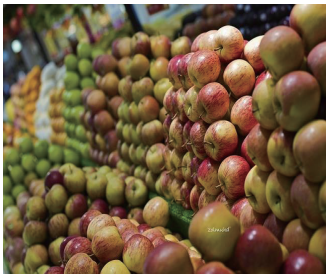
Context: Two men are standing on top of a zebra.
Question: Where would you find this animal in the wild?
Knowledge: In the context given, a zebra is a type of animal. It is a large, striped mammal that lives in Africa. Zebras are xxx
Answer: Africa

(a)



Context: A black and white cat laying on top of a pile of clothes.
Question: Is that a panda or a cat?
Knowledge: In this context, black and white refers to the color of the cat xxx
Answer: cat

(b)



Context: A large display of apples in a grocery store.
Question: What its this a fruit or vegetable?
Knowledge: In the context given, fruit refers to a type of plant that produces seeds and fleshy pulp. xxx
Answer: fruits

(c)

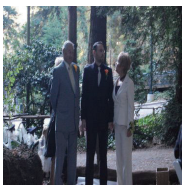
Fig. 6 Impact of background knowledge on inference. Figures 6a and 6b are successful cases and Fig. 6c is a failed case. In Figs. 6a and 6b, the correct answer is hit in the background knowledge, which helps the model predict the correct answer. In Fig. 6c, the correct answer is hit in background knowledge, but the model makes the failed prediction.

Context: A group of men in suits standing in front of a robot.
Question: What type of clothing are the men wearing?
Knowledge: xxx
Answer: suit

Example 1
Context: A man in a suit sits inside of a car.
Question: What type of outfit is this man wearing?
Knowledge: xxx
Answer: suit

Example 2
Context: A smiling business woman is standing next to a business man.
Question: What type of attire is this?
Knowledge: xxx
Answer: business

Example 3
Context: A group of business men in suits and ties talking to kids.
Question: What world describes the men's attire?
Knowledge: xxx
Answer: dress



Example 4
Context: Two men in suits standing next to a women.
Question: What type of outfit is the woman wearing?
Knowledge: xxx
Answer: suit

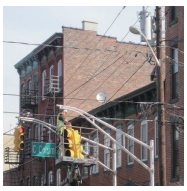
(a)

Context: A fire place sitting in a living room next to a window.
Question: What is this fireplace made of?
Knowledge: xxx
Answer: brick

Example 1
Context: Freight train going over a bridge on a river.
Question: What is this structure made of?
Knowledge: xxx
Answer: steel

Example 2
Context: Three images of the process of a bathroom remodel.
Question: What material is the bathtub made out of?
Knowledge: xxx
Answer: ceramic

Example 3
Context: Vehicles drive past an obelisk topped by a clock.
Question: What is the clock tower made out of?
Knowledge: xxx
Answer: brick



Example 4
Context: An intersection containing a large brick building, traffic lights, and a street sign.
Question: What are the buildings made of?
Knowledge: xxx
Answer: brick

(b)

Fig. 7 Impact of in-content examples on inference. Because in-context examples play a key role in the in-context learning paradigm of LLMs, we show cases where in-context examples help the model predict. The red font indicates that the correct answer is hit in the in-context examples.

Please answer the question according to the context and knowledge.
======
Context: A yellow diamond shaped road sign on the right side of a road.
Question: What does the sign indicate?
Knowledge: A road sign is a sign that is placed on a road to provide information to drivers. They are often used to indicate speed limits, road conditions, and other information. xxxxxx
Answer: roundabout
===
Context: A black and white shot of two people on their motorcycle.
Question: Is which type of road are the people riding?
Knowledge: In this context, people are the two people riding the motorcycle. A motorcycle is a two-wheeled vehicle that is powered by anengine. It is a popular form of transportation in many countries. In this context, the term “road” refers to a paved surface that is used for driving. xxxxxx
Answer: highway
===
Context: A car is parked a the end of a wooded street.
Question: What traffic sign is backwards?
Knowledge: A car is a type of vehicle. It is a four-wheeled motor vehicle that is powered by an internal combustion engine. Cars are typically used for transportation, but can also be used for recreation or racing. A traffic sign is a sign that is used to communicate information to drivers. xxxxxx
Answer: yield
===
Context: Many different cars driving down a city road.
Question: What kind of road is this?
Knowledge: Cars are vehicles that are driven on roads. They are a common form of transportation in urban areas. A road is a path or route that is used for traveling. It can be paved or unpaved, and can be used for a variety of purposes, including transportation, recreation, or agriculture. xxxxxx
Answer: highway
======
Context: A street sign is shown with a tree in the background.
Question: Where does the word shown here come from?
Knowledge: A street sign is a sign that provides information about a street or road. It can be used to provide directions, identify the name of the street, and more. A tree is a woody plant that grows in the ground. Trees can be found in many different environments, including forests, parks, and backyards. xxxxxx
Answer:

Fig. 8 Prompt consisting of three parts: (1) The prompt head is used to describe the task; (2) Some in-context examples are used for LLM learning; (3) Test input is in the same format as the in-context example, but the answer is left blank for LLM prediction. We show the prompt with four in-context examples for better visualization.

number of in-context examples. It is evident that as the number of in-context examples increases, the inference time also extends. This correlation is attributed to the augmented length of the context, leading to longer processing time for the model.

4.8 Evaluation on additional datasets

To further validate the robustness of our method, we conduct additional evaluations on the DAQUAR dataset. To facilitate evaluation, we select one-tenth of the test set as the evaluation set to report the results. We compare with the best baseline model, Prophet, and show comparisons under different in-context example settings. Figure 10 illustrates the results. We find that our method outperforms the best baseline model under different parameter settings. The results again demonstrate the robustness of our method.

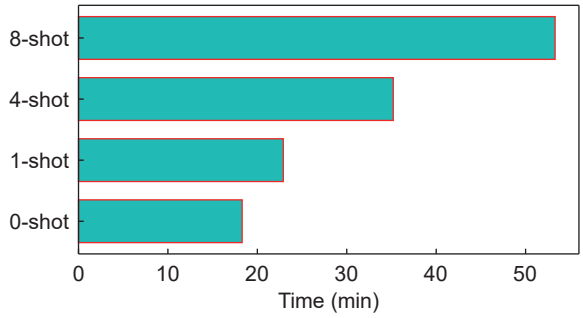


Fig. 9 Inference time required for different numbers of in-context examples.

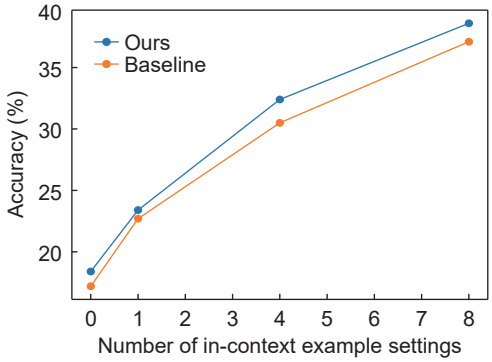


Fig. 10 Comparison on DAQUAR dataset. On the DAQUAR dataset, we compare against the best baseline model under different numbers of in-context example settings.

5 Conclusion

We introduce PLLMKI, a knowledge-based VQA framework utilizing the knowledge-enhanced LLM. Unlike previous works, PLLMKI leverages the LLM to generate background knowledge, integrating it into another LLM for inference, as opposed to using knowledge graphs to enhance the LLM. Our framework consists of three components: (1) A vanilla VQA model is used to get in-context examples, and a captioning model is adopted to convert image-question-answer to context-question-answer; (2) LLM_1 is used to generate background knowledge, and then the knowledge is integrated into the prompt, i.e., context-question-knowledge-answer; (3) The prompt is input into LLM_2 to predict the answer. It is noteworthy that our framework relies solely on open LLMs, incurring no additional costs. Experiments demonstrate the superior performance of our approach on two knowledge-based VQA datasets, OK-VQA and A-OKVQA. As a next step, we consider applying the agent concept to knowledge-based VQA by designing a variety of tools for knowledge enhancement, including retrieval from knowledge graphs and knowledge bases, generation using LLMs, and so on. The agent technology is then utilised to automate the planning in order to inject knowledge in a flexible manner.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62272100), Consulting Project of Chinese Academy of Engineering (No. 2023-XY-09), and Fundamental Research Funds for the Central Universities.

References

[1] Q. Wu, P. Wang, X. Wang, X. He, and W. Zhu, Knowledge-based VQA, in *Visual Question Answering*, Q. Wu, P. Wang, X. Wang, X. He, and W. Zhu, eds. Singapore: Springer, 2022, pp. 73–90.

[2] S. Manmadhan and B. C. Kovoor, Visual question answering: a state-of-the-art review, *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5705–5745, 2020.

[3] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, FVQA: Fact-based visual question answering, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, 2018.

[4] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, OK-VQA: A visual question answering benchmark requiring external knowledge, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3190–3199.

[5] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, A-OKVQA: A benchmark for visual question answering using world knowledge, in *European Conference on Computer Vision*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds. Cham, Switzerland: Springer, 2022, pp. 146–162.

[6] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 14106–14116.

[7] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, An empirical study of GPT-3 for few-shot knowledge-based VQA, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, pp. 3081–3089, 2022.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv: 2005.14165, 2020.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-

intensive NLP tasks, in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds. New York, NY, USA: Curran Associates, Inc., 2020, pp. 9459–9474.

[10] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, et al., SpectralGPT: Spectral remote sensing foundation model, *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2024.3362475.

[11] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks, *Remote. Sens. Environ.*, vol. 299, p. 113856, 2023.

[12] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, An augmented linear mixing model to address spectral variability for hyperspectral unmixing, *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, 2019.

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, VQA: Visual question answering, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Santiago, Chile, 2015, pp. 2425–2433.

[14] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, Visual question answering using deep learning: A survey and performance analysis, in *Computer Vision and Image Processing*, S. K. Singh, P. Roy, B. Raman, and P. Nagabhushan, eds. Singapore: Springer, 2021, pp. 75–86.

[15] P. Sun, W. Zhang, S. Li, Y. Guo, C. Song, and X. Li, Learnable depth-sensitive attention for deep RGB-D saliency detection with multi-modal fusion architecture search, *Int. J. Comput. Vis.*, vol. 130, no. 11, pp. 2822–2841, 2022.

[16] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, Multi-modal 3D object detection in autonomous driving: A survey, *Int. J. Comput. Vis.*, vol. 131, no. 8, pp. 2122–2152, 2023.

[17] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, In defense of grid features for visual question answering, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 10264–10273.

[18] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, VinVL: Revisiting visual representations in vision-language models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 5575–5584.

[19] L. Li, Z. Gan, Y. Cheng, and J. Liu, Relation-aware graph attention network for visual question answering, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 10312–10321.

[20] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, Deep multimodal neural architecture search, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 3743–3752.

[21] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, Deep modular co-attention networks for visual question answering, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 6274–6283.

[22] Y. Cui, Z. Yu, C. Wang, Z. Zhao, J. Zhang, M. Wang, and J. Yu, ROSITA: Enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, 2021, pp. 797–806.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, arXiv preprint arXiv: 2201.12086, 2022.

[24] M. Zhou, L. Yu, A. Singh, M. Wang, Z. Yu, and N. Zhang, Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 16464–16473.

[25] M. Malinowski, M. Rohrbach, and M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Santiago, Chile, 2015, pp. 1–9.

[26] Z. Yu, J. Yu, J. Fan, and D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 1839–1848.

[27] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, Neural module networks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 39–48.

[28] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, Multi-modal answer validation for knowledge-based VQA, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, pp. 2712–2721, 2022.

[29] Z. Shao, Z. Yu, M. Wang, and J. Yu, Prompting large language models with answer heuristics for knowledge-based visual question answering, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 14974–14983.

[30] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Trans. Knowl. Data Eng.*, doi: 10.1109/TKDE.2024.3352100.

[31] X. Zou, A survey on application of knowledge graph, *J. Phys.*: *Conf. Ser.*, vol. 1487, no. 1, p. 012016, 2020.

[32] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, ERNIE: Enhanced language representation with informative entities, in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 1441–1451.

[33] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, et al., ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation, arXiv preprint arXiv: 2107.02137, 2021.

[34] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, What makes good in-context examples for GPT-3?

in *Proc. Deep Learning Inside Out* (*DeeLIO 2022*): *The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, Dublin, Ireland, 2022, pp. 100–114.

[35] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, MUTAN: Multimodal tucker fusion for visual question answering, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2631–2639.

[36] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu, Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering, in *Proc. 29th Int. Joint Conf. Artificial Intelligence*, Yokohama, Japan, 2020, pp. 1097–1103.

[37] F. Gardères, M. Ziaeefard, B. Abeloos, and F. Lecue, ConceptBert: Concept-aware representation for visual question answering, in *Proc. Findings of the Association for Computational Linguistics*: *EMNLP 2020*, Virtual Event, 2020, pp. 489–498.

[38] M. Luo, Y. Zeng, P. Banerjee, and C. Baral, Weakly-supervised visual-retriever-reader for knowledge-based question answering, in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021, pp. 6417–6431.

[39] F. Gao, Q. Ping, G. Thattai, A. Reganti, Y. N. Wu, and P. Natarajan, Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 5057–5067.

[40] Y. Guo, L. Nie, Y. Wong, Y. Liu, Z. Cheng, and M. Kankanhalli, A unified end-to-end retriever-reader framework for knowledge-based VQA, in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 2061–2069.

[41] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, Pythia v0.1: The winning entry to the VQA challenge 2018, arXiv preprint arXiv: 1807.09956, 2018.

[42] J. Lu, D. Batra, D. Parikh, and S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. New York, NY, USA: Curran Associates, Inc., 2019, pp. 13–23.

[43] R. Mokady, A. Hertz, and A. H. Bermano, Clipcap: Clip prefix for image captioning, arXiv preprint arXiv: 2111.09734, 2021.

[44] H. Tan and M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, China, 2019, pp. 5100–5111.

[45] A. Kamath, C. Clark, T. Gupta, E. Kolve, D. Hoiem, and A. Kembhavi, Webly supervised concept expansion for general purpose vision models, in *Computer Vision—ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds. Cham, Switzerland: Springer, 2022, pp. 662–681.

[46] S. Ravi, A. Chinchure, L. Sigal, R. Liao, and V. Shwartz, VLC-BERT: Visual question answering with contextualized commonsense knowledge, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision* (*WACV*), Waikoloa, HI, USA, 2023, pp. 1155–1165.

[47] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in visual question answering, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 6325–6334.

[48] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2017.

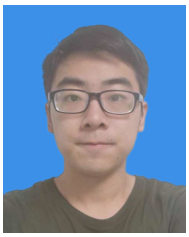
[49] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv: 2302.13971, 2023.

[50] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv: 2307.09288, 2023.

[51] F. Ilievski, P. Szekely, and B. Zhang, CSKG: The commonsense knowledge graph, in *The Semantic Web*, R. Verborgh, K. Hose, H. Paulheim, P. Champin, M. Maleshkova, O. Corcho, P. Ristoski, and M. Alam, eds. Cham, Switzerland, Springer, 2021, pp. 680–696.

Zhongjian Hu is currently pursuing the PhD degree at the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include artificial intelligence, natural language processing, etc.

Peng Yang is a professor at the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include artificial intelligence, natural language processing, big data, etc.



Fengyuan Liu is currently pursuing the master degree at the Southeast University - Monash University Joint Graduate School (Suzhou), Southeast University, China. His research interests include artificial intelligence, large language models, etc.

Xingyu Liu is currently pursuing the master degree at the Southeast University - Monash University Joint Graduate School (Suzhou), Southeast University, China. Her research interests include AI agents, large language models, etc.

Yuan Meng is currently pursuing the PhD degree at the School of Computer Science and Engineering, Southeast University, Nanjing, China. Her research interests include knowledge graphs, natural language processing, etc.