Detecting Knowledge Boundary of Vision Large Language Models by Sampling-Based Inference

Anonymous ACL submission

Abstract

Despite the advancements made in Visual Large Language Models (VLLMs), like text Large Language Models (LLMs), they have limitations in addressing questions that require real-time information or are knowledge-intensive. Indiscriminately adopting Retrieval Augmented Generation (RAG) techniques is an effective yet expensive way to enable models to answer queries beyond their knowledge scopes. To mitigate the dependence on retrieval and simultaneously maintain, or even improve, the performance benefits provided by retrieval, we propose a method to detect the knowledge boundary of VLLMs, allowing for more efficient use of techniques like RAG. Specifically, we propose a method with two variants that fine-tunes a VLLM on an automatically constructed dataset for boundary identification. Experimental results on various types of Visual Question Answering datasets show that our method successfully depicts a VLLM's knowledge boundary based on which we are able to reduce indiscriminate retrieval while maintaining or improving the performance. In addition, we show that the knowledge boundary identified by our method for one VLLM can be used as a surrogate boundary for other VLLMs. Code will be released at `https://code.github.com`

1 Introduction

The great advancements in language models have led to the integration of image encoding and understanding capabilities (Achiam et al., 2023; Lu et al., 2024a; Wang et al., 2024), significantly enhancing the performance of a series of pre-trained Visual Large Language Models (VLLMs) in tasks involving Visual Question Answering (VQA). Despite these advancements, akin to text Large Language Models (LLMs) (Touvron et al., 2023; Workshop et al., 2022; Brown et al., 2020; Zhang et al., 2024b), VLLMs remain constrained by the boundaries of their knowledge (Lin and Byrne, 2022). As a result, their ability to accurately respond to content outside the model's knowledge scope, such as knowledge-intensive questions, real-time news, and queries with dynamic answers, is considerably limited.

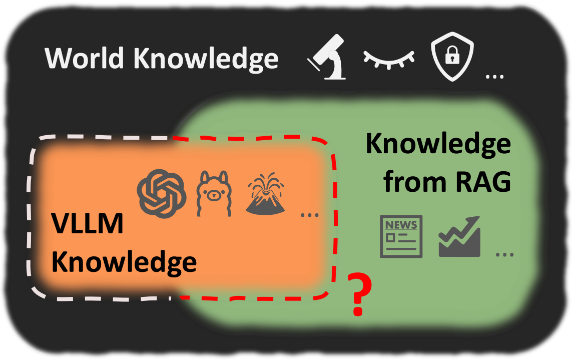

Figure 1: VLLMs Knowledge Boundary concept. The black part represents all the knowledge humans have explored, and the orange and green parts represent knowledge possessed by VLLMs and knowledge that can be retrieved from external sources respectively. They overlap in some areas and the boundary between them remains unclear. The overall knowledge boundary of VLLMs can be differentiated into two parts that overlap with knowledge between RAG and world knowledge. Our method aims to identify both, and we conduct experiments to validate the potential VQA performance improvements using RAG.

Some works study the knowledge boundary of text LLMs (Anonymous, 2024; Cheng et al., 2024; Zhang et al., 2024b; Ren et al., 2023) via prompt-based or SFT-based methods. As of yet, there has been limited research on the methodology for determining the knowledge boundaries of VLLMs. In practical applications, to answer VQA queries outside its knowledge scope, indiscriminately employing Retrieval Augmented Generation (RAG) techniques is often a viable solution. Although this approach has been proven to enhance the (V)LLMs' performance (Wang et al., 2021; Lewis et al., 2020; Chen et al., 2017), the comprehensive reliance on retrieval methods incurs significant latency due to the retrieval steps and the introduction of excessively long inputs. (Chevalier et al., 2023; Zhang et al., 2024a; Chen et al., 2024)

To mitigate the dependence on retrieval for answering questions and simultaneously maintain, or improve, the performance benefits provided by retrieval, we aim to develop a method that can depict the knowledge boundary of a VLLM. In this paper, we employ a method with two variants to delineate the knowledge boundaries of a VLLM by fine-tuning a VLLM on data constructed based on sampling the responses of the VLLM.

With the ability to depict the knowledge boundary of a VLLM, we then adopt RAG techniques to validate the accuracy of the identified boundary in various held-out datasets. We conduct experiments using a variety of VQA datasets, including three knowledge-intensive datasets, two non-knowledge-intensive datasets, and one mixed dataset. After determining whether a query falls within the knowledge boundary, we use RAG to assess the potential improvements the retrieved information provides to the queries falling out of the knowledge boundary. Our experimental results reveal that on a mixed dataset, which contains both non-knowledge-intensive and knowledge-intensive queries simulating real situations, our method outperforms the indiscriminative use of RAG (denoted "All RAG") and prompt-based baseline with 50.67% retrieval reduction. The fine-tuned knowledge boundary model lowers the retrieving ratio on less knowledge-intensive data and obtains close or even better performance compared to the "All RAG" setting. Besides, we show that the fine-tuned VLLM for boundary identification for one VLLM can be used as a surrogate boundary identifier for other VLLMs.

To sum up, our contributions are as follows:

1. We propose a method with two variants that detects the knowledge boundary of a VLLM.
2. Experimental results show that we maintain, or even improve, the performance of the VLLM on various types of data while lowering the ratio of using RAG, and our method outperforms the "All RAG" setting and other baselines on a dataset simulating real situations.
3. We show that the knowledge boundary for one VLLM can be used as a surrogate boundary for other VLLMs, to reduce retrieval while maintaining or improving the performance.

2 Method

We propose a method with two variants that fine-tunes a VLLM, which can depict the *hard* or *soft* knowledge boundary of VLLMs. The proposed method relies only on (V)LLMs and does not require manual annotation. In the following sections, we first introduce the background and necessary notations. Then we give details on constructing two types of datasets for fine-tuning a VLLM for knowledge boundary approximation.

2.1 Background

Consider a Visual Question Answering query $\boldsymbol{q}$ with gold text answer $\boldsymbol{a}$, where $\boldsymbol{q}$ contains image(s) $\boldsymbol{q_i}$ and a text query $\boldsymbol{q_t}$. Also, contexts $\boldsymbol{k}$ related to $\boldsymbol{q}$ can be retrieved from a given corpus, where $\boldsymbol{k}$ can refer to the collection of both texts and images. Given a VL model, parameterized by θ, we can answer the query with or without RAG by running decoding (Dec) on the model:

$$\begin{aligned} \boldsymbol{y_n} &= Dec_\theta(\boldsymbol{y}|\boldsymbol{q}) \\ \boldsymbol{y_r} &= Dec_\theta(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{k}) \end{aligned} \tag{1}$$

where $\boldsymbol{k}$ might also contain prompts connecting related content and it is omitted here for simplicity.

It is acknowledged that VLLMs have a limited knowledge scope (Lin and Byrne, 2022; Wu et al., 2022), denoted as S, and the boundary is a rather vague concept and is hard to depict accurately.

2.2 Sampling

To approximate whether a query $\boldsymbol{q}$ should lie in VLLMs' knowledge scope S, we run repeated sampling of a VLLM and collect its outputs. The sampling methods include but are not limited to, top-p sampling and top-k sampling. By running R times sampling, we obtain R outputs given query $\boldsymbol{q}$:

$$\boldsymbol{y^{(i)}} = Dec_\theta(\boldsymbol{y}|\boldsymbol{q}), i \in \{1, 2, ..., R\} \tag{2}$$

After obtaining the R predictions, a text LLM is prompted[1] to evaluate each prediction $y^{(i)}$ where the gold answer is also given. Subsequently a score $s_i \in [s_w, s_c]$ is provided by this text LLM. We define the score range within s_w and s_c, where s_c indicates a perfectly correct answer and s_w indicates a wrong answer. Then an average score is calculated over R scores, indicating the overall

[1] The prompt is referenced from Liu (2022). Please refer to our code for a detailed definition.

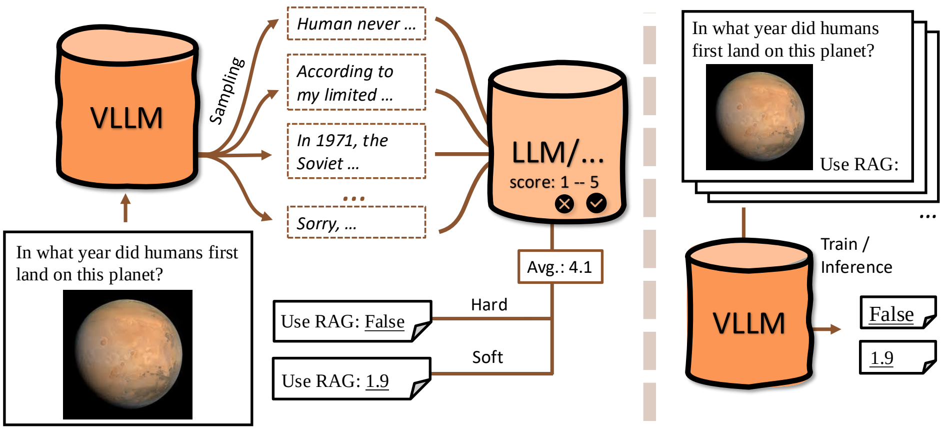

Figure 2: Method illustration of training a Knowledge Boundary model.

performance of this query:

$$s = mean(s_i), i \in \{1, 2, ..., R\} \tag{3}$$

and we note that s is also $\in [s_w, s_c]$.

2.3 Training

The score s is used to construct the knowledge boundary training data. We differentiate our method into two variants. A VLLM is adopted to train on the knowledge boundary training data. We denote the parameters by ϕ [2].

Hard Knowledge Boundary By setting a threshold ϵ, we deem the queries with score $s \geq \epsilon$ inside the knowledge boundary S and the rest outside S. The query $\boldsymbol{q}$, together with proper prompts P_h, will be constructed into a training sample $\boldsymbol{x}(\boldsymbol{q}, P_h)$ as shown in Sec. A.1. For any $\boldsymbol{x}(\boldsymbol{q}, P_h)$ in the training dataset, we define the training objective J_h w.r.t. ϕ as follows:

$$\begin{aligned} J_h(\phi) = &- \sum_{\boldsymbol{x}(\boldsymbol{q}, P_h): \boldsymbol{q} \notin S} \log P_\phi(\text{“True”} | \boldsymbol{x}(\boldsymbol{q}, P_h)) \\ &- \sum_{\boldsymbol{x}(\boldsymbol{q}, P_h): \boldsymbol{q} \in S} \log P_\phi(\text{“False”} | \boldsymbol{x}(\boldsymbol{q}, P_h)) \end{aligned} \tag{4}$$

where $P_\phi(a|b)$ stands for the probability model ϕ predicts on a given input b. ϕ is optimized by minimizing $J_h(\phi)$.

Soft Knowledge Boundary Setting a threshold to binarily classify the queries might be an overly rigid method and there is no room for adjustment when the knowledge boundary model performs poorly in possibly unseen scenarios unless we adjust ϵ and retrain the model. Thus, we also propose a method that can depict a softer boundary. Recall that for query $\boldsymbol{q}$, the average score s over R model predictions ranges in $[s_w, s_c]$, where s_w indicates a wrong answer and s_c indicates a correct one. We linearly flip the score, for example, the new score $s^{'} = s_w$ represents a strong tendency for external knowledge while $s^{'} = s_c$ represents a refusal to external knowledge.

The query $\boldsymbol{q}$, together with prompts P_s, will be constructed into a training sample $\boldsymbol{x}(\boldsymbol{q}, P_s)$ as shown in Sec. A.1. For any $\boldsymbol{x}(\boldsymbol{q}, P_s)$ in the training dataset, we define the training objective as follows:

$$J_s(\phi) = - \sum_{\boldsymbol{x}(\boldsymbol{q}, P_s)} \log P_\phi(s^{'} | \boldsymbol{x}(\boldsymbol{q}, P_s)) \tag{5}$$

where ϕ is optimized by minimizing $J_s(\phi)$.

By optimizing objective 4, we get a Hard Knowledge Boundary model HKB_ϕ that can take a VQA sample and predict a *binary* output “True” or “False” indicating whether the RAG technique can help solve this query. Similarly, a Soft Knowledge Boundary model SKB_ϕ that can predict a *soft score*, ranging from s_w to s_c, is trained by optimizing objective 5:

$$\begin{aligned} HKB_\phi(\boldsymbol{x}(\boldsymbol{q}, P_h)) &= \text{True / False} \\ SKB_\phi(\boldsymbol{x}(\boldsymbol{q}, P_s)) &\in [s_w, s_c] \end{aligned} \tag{6}$$

2.4 Application of RAG in Our Method

An indicator function is defined to map the prediction of a Hard/Soft Knowledge Boundary model to

[2] Our method does not require ϕ to be identical to the parameters of the previously sampled VLLM.

a real search decision:

$$\mathbb{I}(\boldsymbol{q},\boldsymbol{k}) = \begin{cases} \boldsymbol{k}, \text{if } HKB_{\phi}(\boldsymbol{x}(\boldsymbol{q},P_h)) == \text{true} \\ \quad \text{or } SKB_{\phi}(\boldsymbol{x}(\boldsymbol{q},P_s)) \geq \epsilon \\ None, \text{else} \end{cases} \tag{7}$$

Then we can combine the decoding with or without RAG stated in equation 1 into:

$$\boldsymbol{y_{kb}} = Dec_{\theta,\phi}(\boldsymbol{y}|\boldsymbol{q}, \mathbb{I}(\boldsymbol{q},\boldsymbol{k})) \tag{8}$$

3 Experiment

3.1 Setup

3.1.1 Training Data

With method stated in Sec. 2.2 and 2.3, we adopt InfoSeek (Chen et al., 2023), OK-VQA (Marino et al., 2019), VQAv2.0 (Goyal et al., 2017), MMBench (Liu et al., 2025), and MME (Fu et al., 2023) to construct the training set where we randomly sample two subsets from InfoSeek and VQAv2.0 respectively due to their large sizes. Table 1 presents the detailed sizes for each dataset we use along with the average scores s. In our experiments $s_w = 1$ and $s_c = 5$. We adopt all these datasets to increase the diversity of queries as much as possible. Below is a brief description of each dataset.

InfoSeek is designed to assess the capability of models to seek and incorporate external information for question answering. It features a variety of queries that necessitate fact retrieval and reasoning that go beyond the provided context.

OK-VQA is a dataset where images are paired with open-ended questions that require answers stemming from general knowledge that extends beyond the image alone.

VQAv2.0 is a comprehensive VQA dataset that requires interpretation or understanding of the visual content. It features a diverse and balanced range of answers.

MMBench is a benchmarking suite for evaluating multi-modal understanding, ensuring that multi-modal machine learning systems can effectively process and synthesize data from different sources.

MME is focused on tasks related to multi-modal entity recognition and extraction. The dataset contains annotations of text and images with multi-modal entities that need to be identified or linked.

Source	# Samples	Avg. Score ± std.
InfoSeek	216000	1.82± 1.17
OK-VQA	9009	3.70± 1.48
VQAv2.0	108000	4.27± 1.36
MMBench (en)	4329	3.92± 1.72
MME	2374	4.15± 1.63

Table 1: Training set sources and statistics. Answers are sampled from Qwen-VL-Chat (Bai et al., 2023) and scores are evaluated by Qwen-Max (Team, 2024)

Human-Labeled A group of annotators is asked to annotate whether RAG can help solve a VQA sample. We construct this data to form a reference setting.

3.1.2 Test Data

As we aim to construct a model that can take various input queries and make good judgments about the knowledge boundary, we adopt held-out data to evaluate the final VQA performance. We summarize the overall RAG Effect on each data in Table 2 and a brief introduction as follows.

Life VQA We collect a set of VQA data from people's daily lives and extract the ones current VLLMs do not perform well, which is used to verify whether our model decides to resort to RAG for help. We will release this data along with the code and name it Life VQA.

Private VQA is an internal dataset spanning broad categories, including animals, plants, architecture, geographic locations, etc. Due to the complexity of the backgrounds and the presence of multiple objects, this collection poses a notable challenge for advanced visual reasoning and understanding. This dataset will not be released for now.

Dyn-VQA is released by Li et al. (2024) and contains three types of questions: questions with rapidly changing answers, questions requiring multi-modal knowledge and multi-hop questions. This dataset is a challenging one in our evaluation. *Gold query* is annotated by Li et al. (2024) that combines the text query and image to be used to retrieve useful information.

NoCaps (Agrawal et al., 2019) is an open-domain image captioning dataset derived from Open Images (Krasin et al., 2017), focusing on

Test Data	RAG Effect
Life VQA	High
Private VQA	Medium
Dyn-VQA	High
NoCaps	Low
Visual7W	Low
Mix	?

Table 2: Test data property illustration that whether RAG is helpful in answering the queries.

generating captions for a diverse array of objects and scenes. We sample a subset of size 500.

Visual7W (Zhu et al., 2016) is a VQA dataset containing images from COCO (Lin et al., 2014), paired with seven types of questions (who, what, when, where, how, why and which). It aims to evaluate models' abilities in object recognition and deeper reasoning within visual contexts.

Mix is a composite dataset consisting of 100 samples from each of the aforementioned datasets. It is designed to integrate the characteristics of each dataset and simulate real-world scenarios. Thus the effect of RAG on this dataset is mixed and hard to predict intuitively.

3.1.3 Use of RAG

We aim not only to locate the queries that need RAG to answer better but also adopt retrieval techniques to verify the final VQA performance with the search decision HKB_ϕ and SKB_ϕ defined in equations 6. We note that although there are various options for retrieval, such as text search and image search, we do not design detailed methods to determine the best option in this paper. Instead, we directly use text search (Google) for Dyn-VQA and image search (Bing) for the rest for better retrieval information quality towards answering the question. We note that Dyn-VQA is a challenging dataset that includes many samples of multi-hop property, therefore we use the *golden query* Li et al. (2024) have summarized for retrieving useful information.

In the following sections, the "No RAG" setting refers to the performance of only VLLMs and no retrieval information is given, and "All RAG" refers to always incorporating RAG. "Prompt-based" refers to prompting the model that is sampled to adopt RAG or not.

3.1.4 Base Models

When constructing the training set according to the method stated in Sec. 2.2, we experiment with Qwen-VL-Chat that is used to be sampled $R = 30$ times and fine-tuned according to Sec. 2.3. Refer to Sec. A.2 for detailed training settings. Qwen-Max is prompted to score the R predictions to get scores s_i where we adopt $s_w = 1$ and $s_c = 5$ referenced from Liu (2022).

For Visual Question Answering, we first evaluate the performance of the original Qwen-VL-Chat. In addition, we seek to validate whether the identified knowledge boundary can function as a surrogate boundary for other VLLMs since constructing training datasets through sampling (Sec. 2.3) on (larger) models can be prohibitively expensive. We further validate the surrogate knowledge boundary on the following VLLMs, Deepseek-VL-Chat (Lu et al., 2024b), Qwen-VL-Max (Bai et al., 2023), Qwen-VL-2 (Wang et al., 2024) and GPT4-o (Hurst et al., 2024), to evaluate its potential for generalizing across different VLLMs.

3.2 Main Results

We present our main results of Qwen-VL-Chat in Table 3. Metrics **LLM** represents that the score is evaluated by a text LLM, Qwen-Max, given the model prediction and gold answer. Metrics **Acc.** refers to token accuracy which involves determining the proportion of tokens in the model's predictions that match the tokens in the gold answer. Both Scores range from 0 to 100 and a higher score indicates a higher performance. The % columns refer to the ratio of data that our knowledge boundary model predicts to lie beyond the VLLM's knowledge boundaries. The "Human" column represents the corresponding statistics where the Knowledge Boundary model is trained on the human-labeled data mentioned in Sec. 3.1.1 and we deem it a reference result.

First, the results in the Mix row, which considers all kinds of VQA queries in our setting and simulates a real situation, show that our methods outperform all other baseline and reference settings. Besides, our HKB method lowers the retrieval demand by 23.17% and the SKB method lowers the retrieval demand by 50.67%.

Second, as shown by the % columns and the RAG Effect we summarized in Table 2, our Knowledge Boundary models succeed in predicting a high ratio on test data when RAG can effectively aid in

Dataset	Metric	No RAG	All RAG	Prompt-based	%	HKB	%	SKB	%	Human	%
Life VQA	LLM	30.00	40.70	33.89	12.75%	40.64	96.64%	36.78	61.74%	39.33	71.14%
	Acc.	17.80	36.11	21.38	12.75%	36.11	96.64%	29.44	61.74%	33.36	71.14%
Private VQA	LLM	22.90	24.35	24.95	14.80%	24.50	99.20%	22.89	67.80%	24.20	72.00%
	Acc.	16.26	18.40	17.26	14.80%	18.40	99.20%	17.35	67.80%	18.55	72.00%
Dyn-VQA ch	LLM	19.16	38.95	19.70	6.38%	37.94	95.66%	36.53	84.26%	28.89	46.95%
	Acc.	23.41	43.06	24.37	6.38%	42.71	95.66%	40.97	84.26%	33.13	46.95%
Dyn-VQA en	LLM	21.60	34.93	23.51	14.13%	33.30	89.79%	32.06	76.08%	25.73	29.51%
	Acc.	25.64	41.87	27.58	14.13%	40.66	89.79%	38.51	76.08%	30.83	29.51%
NoCaps	LLM	50.13	30.37	50.13	0.00%	42.50	38.40%	50.13	0.00%	50.13	0.00%
	Acc.	40.50	30.72	40.50	0.00%	36.95	38.40%	40.50	0.00%	40.50	0.00%
Visual7W	LLM	54.48	52.04	55.32	31.36%	52.95	35.37%	54.27	2.96%	54.53	0.52%
	Acc.	44.34	44.94	44.18	31.36%	44.32	35.37%	44.68	2.96%	44.34	0.52%
Mix	LLM	34.44	38.60	34.98	12.67%	39.59	76.83%	**39.93**	49.33%	38.29	38.33%
	Acc.	26.13	32.39	27.23	12.67%	**32.73**	76.83%	30.98	49.33%	31.02	38.33%

Table 3: Main results of Qwen-VL-Chat. Scores are shown in columns except for the % ones. Metrics are evaluated by Qwen-Max (LLM) and Token Accuracy (Acc.). Underlines mark the results that outperform three baseline "No RAG", "All RAG" and "Prompt-based" settings. **Boldface** marks the best results.

answering the query, and it lowers the ratio for data where the queries tend to fall within the knowledge scope of a VLLM.

Third, on the first four datasets where RAG can (greatly) enhance the VQA performance, we show that with our HKB and SKB, the performance is close to that achieved with the "All RAG" setting. For example, with the SKB model, Qwen-VL-Chat archives a 32.06 LLM score on the Dyn-VQA (en) dataset with 76.08% RAG ratio, whereas the "All RAG" setting achieves 34.93. With the HKB model, Qwen-VL-Chat exceeds the "All RAG" setting on Private VQA, even though we note that "All RAG" is a strong setting on this data.

At last, on the NoCaps and Visual7W datasets where VLLMs can perform well without RAG and RAG tends to supply noise, our method can identify a much lower search ratio. Specifically, the search ratio from SKB is close to or equal to zero.

4 Analysis

In this section, we present three analytical experiments. The first one shows the performance of other VLLMs if we employ the identified knowledge boundary as a surrogate. The second shows how the RAG ratio and VQA performance are affected by the threshold defined in the SKB variant. The third one presents the accuracy of VLLM boundary identification on held-in data at training time.

4.1 Surrogate Boundary for Other VLLMs

We assemble around 340 thousand VQA samples from various domains discussed in Sec. 3.1.1. Sampling each data thirty times is prohibitively expensive for closed-source VLLMs. Although different VLLMs possess varying scopes of knowledge intuitively, we believe that it is highly probable that a significant portion of these scopes overlap. For instance, queries regarding recently occurring news events typically fall outside the knowledge boundaries of any model. Thus, we conduct an experiment that validates whether the identified knowledge boundary can function as a surrogate boundary for other VLLMs.

Results are presented in Table 4 and Table 7. From Table 4 Mix row, Qwen-VL-Max, Qwen-VL-2 and GPT4-o achieve better performance than three baseline settings. Deepseek-VL-Chat remains competitive to the "All RAG" setting with LLM metric and outperforms all other settings in Table 7 Mix row. For other datasets, we show that the previously identified knowledge boundary can help maintain the performance with a reduced RAG ratio. For example, GPT4-o achieves 54.83 with only 61.74% RAG ratio while the "All RAG" setting achieves 55.47 on the Life VQA dataset. Deepseek-VL-Chat maintains its performance on the Dyn-VQA (en) dataset compared to the "All RAG" setting and keeps a clear margin compared to the "No RAG" setting with a 23.92% retrieving deduction.

	Metric: LLM	No RAG	All RAG	Prompt- based	%	HKB	%	SKB	%	Human	%
Life VQA	Ds.-VL-Chat	25.54	47.38	27.68	12.75%	46.91	96.64%	41.21	61.74%	41.61	71.14%
	Qwen-VL-Max	43.26	56.38	45.97	12.75%	56.85	96.64%	53.86	61.74%	55.23	71.14%
	Qwen-VL-2	42.55	54.43	46.28	12.75%	54.03	96.64%	52.28	61.74%	53.96	71.14%
	GPT4-o	47.52	55.47	48.26	12.75%	56.14	96.64%	54.83	61.74%	54.90	71.14%
Private VQA	Ds.-VL-Chat	23.01	27.06	23.89	14.80%	26.94	99.20%	26.19	67.80%	25.83	72.00%
	Qwen-VL-Max	35.20	41.90	38.30	14.80%	41.68	99.20%	40.45	67.80%	43.18	72.00%
	Qwen-VL-2	35.16	38.02	36.57	14.80%	37.84	99.20%	35.85	67.80%	38.25	72.00%
	GPT4-o	39.70	38.21	40.06	14.80%	37.85	99.20%	38.83	67.80%	40.21	72.00%
Dyn-VQA ch	Ds.-VL-Chat	21.62	44.10	22.98	6.38%	42.92	95.66%	40.99	84.26%	34.24	46.95%
	Qwen-VL-Max	32.97	51.24	34.23	6.38%	50.86	95.66%	48.24	84.26%	43.33	46.95%
	Qwen-VL-2	32.78	50.74	34.02	6.38%	50.48	95.66%	48.19	84.26%	43.05	46.95%
	GPT4-o	41.91	56.31	42.53	6.38%	56.31	95.66%	54.49	84.26%	48.95	46.95%
Dyn-VQA en	Ds.-VL-Chat	25.58	38.10	27.19	14.13%	36.86	89.79%	36.32	76.08%	29.44	29.51%
	Qwen-VL-Max	37.19	43.98	38.32	14.13%	43.09	89.79%	42.78	76.08%	39.48	29.51%
	Qwen-VL-2	37.12	44.20	37.17	14.13%	42.47	89.79%	42.32	76.08%	40.07	29.51%
	GPT4-o	45.41	50.93	45.24	14.13%	49.88	89.79%	48.75	76.08%	47.14	29.51%
NoCaps	Ds.-VL-Chat	63.67	59.81	63.67	0.00%	61.23	38.40%	63.67	0.00%	63.67	0.00%
	Qwen-VL-Max	62.10	49.66	62.10	0.00%	57.09	38.40%	62.10	0.00%	62.10	0.00%
	Qwen-VL-2	62.10	49.93	62.10	0.00%	56.93	38.40%	62.10	0.00%	62.10	0.00%
	GPT4-o	61.43	63.98	61.43	0.00%	62.12	38.40%	61.43	0.00%	61.43	0.00%
Visual7W	Ds.-VL-Chat	58.34	57.29	57.26	31.36%	57.85	35.37%	58.13	2.96%	58.28	0.52%
	Qwen-VL-Max	58.37	55.51	62.11	31.36%	57.10	35.37%	58.25	2.96%	58.30	0.52%
	Qwen-VL-2	58.16	54.41	62.19	31.36%	56.66	35.37%	57.85	2.96%	58.02	0.52%
	GPT4-o	52.96	47.06	51.82	31.36%	50.87	35.37%	52.89	2.96%	52.87	0.52%
Mix	Ds.-VL-Chat	34.96	45.18	35.71	12.67%	45.08	76.83%	43.35	49.33%	42.20	38.33%
	Qwen-VL-Max	46.54	49.26	47.30	12.67%	50.64	76.83%	51.06	49.33%	52.05	38.33%
	Qwen-VL-2	46.36	47.89	47.46	12.67%	49.31	76.83%	49.29	49.33%	51.41	38.33%
	GPT4-o	51.44	52.90	50.57	12.67%	54.10	76.83%	52.97	49.33%	55.27	38.33%

Table 4: Knowledge Boundary model as a surrogate boundary identifier for other VLLMs.

4.2 Effect of ϵ for SKB

In Sec 3.2, we show the result of the SKB method with the least RAG ratio, i.e., ϵ is set to maintain a low tendency to resort to RAG. Here we show how the overall VQA performance is affected by ϵ. The results of three datasets are illustrated in Fig. 3. The leftmost point of the horizontal axis corresponds to the "All RAG" setting (with $\epsilon = s_w$), while the rightmost point represents the minimal search ratio. Light-coloured dashed lines depict the "No RAG" setting. For the left two data in Fig 3, where RAG can greatly affect the performance, our methods can maintain a clear margin between the "No RAG" setting and obtain a relatively stable performance with a decreased search ratio. For the Mix data where all types of data are fused, our methods can still lower the search ratio while maintaining, or improving, the performance.

4.3 Performance of Knowledge Boundary Identification on Held-In Data

The training results of the Knowledge Boundary model are shown in Table 5. We show that by training Qwen-VL-Chat, it succeeds in modeling the knowledge boundary on held-in data we constructed according to Sec. 2.3.

	Human-labeled	Hard	Soft
Train	96.25	90.50	88.41
Val.	-	91.16	88.96

Table 5: Training and validation results on the held-in dataset. Metrics are shown in the accuracy of "True/False" or score $s^{'}$ prediction. We have a limited number of human-labeled samples thus we do not set a validation set for "Human-labeled" setting.

5 Related Work

5.1 Knowledge Boundary Study of Text LLM

As the LLMs are applied to a wider range of fields, users expect them to perform well on any query. However, inevitably, the knowledge embedded within LLMs does not automatically update over time, resulting in certain queries consistently falling outside the model's knowledge boundaries. Some works study the Knowledge Bound-

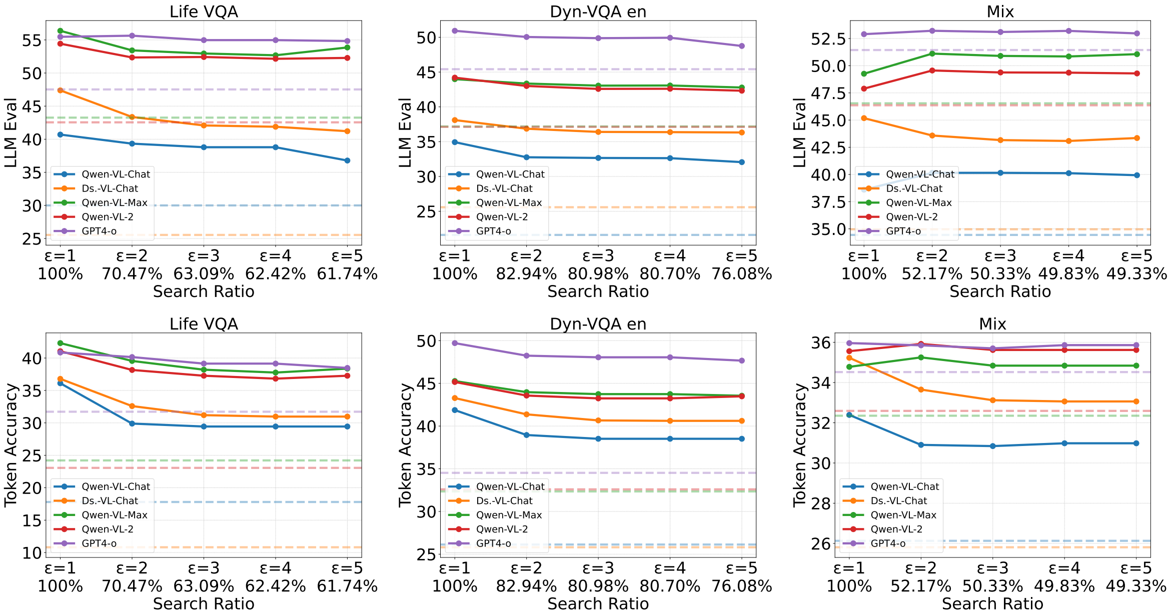

Figure 3: Effect of ϵ. The lighter dashed lines accordingly indicate the performance under each base model's "No RAG" setting.

aries of text LLMs. A commonly used approach prompts LLMs to output content like *"I don't know"* (Anonymous, 2024; Cheng et al., 2024; Ren et al., 2023). Alternatively, another approach is to construct a dataset and perform Supervised Fine-Tuning (SFT) (Zhang et al., 2024b; Cheng et al., 2024; Anonymous, 2024). Both aforementioned types of approaches focus on making the models express *"I know"* or *"I don't know"*. Most aforementioned works find that prompt-based methods are poorly performed.

We contend that this task is actually challenging for two primary reasons. First, regarding whether a model can itself articulate its own knowledge boundaries, considerable debate persists in current research. For example, Ren et al. (2023) states that LLMs struggle to perceive their factual knowledge boundary, and tend to be overconfident, however, Cheng et al. (2024) conclude that the AI assistant can, to a significant extent, identify what it does not know. Second, it is difficult to verify the accuracy of the predicted boundaries.

5.2 Retrieval-Augmented Generation

The RAG technique is widely adopted to help models answer certain queries needing external information in both texts (Jeong et al., 2024; Chen et al., 2024; Lewis et al., 2020) and image-text scenarios (Lin and Byrne, 2022; Wu et al., 2022). However, current RAG techniques are far from being perfect for enhancing (V)LLMs in all settings. For example, Zhang et al. (2024b) finds that for math reasoning and code questions, RAG usually brings noise rather than useful information, and thus RAG may even yield adverse effects. Therefore, more effective utilization of RAG can not only result in savings of time and computational resources but also enhance performance in certain scenarios.

6 Conclusion

In this paper, we introduce a method with two variants that fine-tunes a VLLM on an automatically constructed dataset for boundary identification. This method mitigates the reliance on RAG techniques, which introduce significant latency and long input sequences. Our experiments across diverse held-out VQA datasets, including knowledge-intensive, non-knowledge-intensive, and mixed datasets, demonstrate that our method not only maintains or enhances VLLM performance but also lowers the RAG ratio. Additionally, the fine-tuned knowledge boundary exhibits versatility by functioning as a surrogate for other VLLM series, facilitating retrieval reduction without compromising performance. These findings underscore the efficacy of our approach in optimizing the balance between retrieval dependence and model performance, paving the way for more efficient and effective deployment of VLLMs in practical applications.

7 Limitations

In this paper, we do not design detailed methods to distinguish the search type, such as text search and image search, towards answering a VQA sample. Experiments utilizing training data sampled from larger VLLMs are currently lacking. Both limitations will be addressed in our future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Anonymous. 2024. Refine knowledge of large language models via adaptive contrastive learning. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Zhuo Chen, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Kewei Tu. 2024. Improving retrieval augmented open-domain question-answering with vectorized contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7683–7694, Bangkok, Thailand. Association for Computational Linguistics.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. *arXiv preprint 2305.14788*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S. Yu, Fei Huang, and Jingren Zhou. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jerry Liu. 2022. LlamaIndex.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024b. Deepseek-vl: Towards real-world vision-language understanding.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Qwen Team. 2024. Introducing qwen1.5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024a. Long context compression with activation beacon.

Zhen Zhang, Xinyu Wang, Yong Jiang, Zhuo Chen, Feiteng Mu, Mengting Hu, Pengjun Xie, and Fei Huang. 2024b. Exploring knowledge boundaries in large language models for retrieval judgment. *arXiv preprint arXiv:2411.06207*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Appendix

A.1 Training Examples

Hard Knowledge Boundary Query $\boldsymbol{q}$, together with prompts P_h (in blue)[3], will be constructed into a training sample $\boldsymbol{x}(\boldsymbol{q}, P_h)$ as follows:

> You are an assistant capable of deciding whether a search is needed in a multimodal question-answering scenario. Below, I will provide you with a multimodal question that includes a text question and an image link. Please respond with "true" or "false," indicating whether a search is necessary (true) or not (false) to answer this multimodal question.
> <ST_1>
> Text question: $\boldsymbol{q_t}$
> <Image>: $\boldsymbol{q_i}$
> <ST_2>

Soft Knowledge Boundary Query $\boldsymbol{q}$, together with prompts P_s (in blue), will be constructed into a training sample $\boldsymbol{x}(\boldsymbol{q}, P_s)$ as follows:

> You are an assistant capable of deciding whether a search is needed in a multimodal question-answering scenario. Below, I will provide you with a multimodal question that includes a text question and an image link. Please respond with a score ranging from 1.0 to 5.0 indicating whether a search is necessary or not to answer this multimodal question.
>
> Follow these guidelines for scoring:
> - Your score has to be between 1.0 and 5.0, where 1.0 stands for an unnecessary search and 5.0 stands for a necessary search.
> - The score does not have to be integer.
> Example Response:
> 4.0
>
> <ST_1>
> Text question: $\boldsymbol{q_t}$
> <Image>: $\boldsymbol{q_i}$
> <ST_2>
> Your score:

Base Model	Qwen-VL-Chat
LoRA	Q, K, V
LoRA Rank	8
LoRA Alpha	32
Learning Rate	1e-4
Optimizer	AdamW
LR Scheduler	Linear
Precision	bf16
Batch Size	1
GPU	NVIDIA A100-SXM4-80GB

Table 6: Detailed hyperparameters.

A.2 Training Details and Hyperparameters

Recall that our methods need to train a VLLM, parameterized by ϕ, as a Knowledge Boundary model discussed in Sec. 2.3. In experiments, we adopt LoRA (Hu et al., 2021) to optimize ϕ and the related hyperparameters are shown in Table 6.

A.3 Supplementary Results of Token Accuracy

We provide the supplementary experimental results for Sec 4.1 where the token accuracy metrics are shown in Table 7. Similar conclusions can be made.

A.4 Supplementary Results on MMMU Dataset

In this section, we show the experimental results of our methods on a challenging dataset, MMMU[4] (Yue et al., 2024) in Table 8. MMMU is a dataset containing VQA samples demanding college-level subject knowledge and deliberate reasoning, and it is hard to verify the knowledge boundary that our methods depict by simply adopting RAG.

The results in Table 8 show the Knowledge Boundary model trained by human-labeled data helps achieve the best performance. It verifies that the aforementioned Human-labeled training data is effective. In addition, we show that our methods also exhibit substantial potential within this setting, in which both the HKB and SKB models predict a high search ratio over MMMU. We contend that the suboptimal performance of this dataset arises because it lies beyond the knowledge boundaries, that are challenging to validate using RAG, as delineated by the white dashed lines in Fig. 1. We

[3] where <ST_*> means optional special tokens to specify the position of $\boldsymbol{q}$ and indicate the output starting position after <ST_2>. The detailed format of <ST_*> and <Image> tokens might need to be modified according to different VL model input formats.

[4] We converted the dataset's original multiple-choice format into a conventional VQA format to ensure consistency with the aforementioned experimental settings.

	Metric: Acc.	No RAG	All RAG	Prompt-based	%	HKB	%	SKB	%	Human	%
Life VQA	Ds.-VL-Chat	10.82	36.79	14.12	12.75%	36.12	96.64%	30.97	61.74%	30.50	71.14%
	Qwen-VL-Max	24.21	42.30	27.66	12.75%	41.96	96.64%	38.37	61.74%	38.20	71.14%
	Qwen-VL-2	23.06	41.05	27.64	12.75%	40.71	96.64%	37.27	61.74%	37.05	71.14%
	GPT4-o	31.72	40.85	32.81	12.75%	40.85	96.64%	38.47	61.74%	41.88	71.14%
Private VQA	Ds.-VL-Chat	15.51	19.75	16.65	14.80%	19.75	99.20%	18.20	67.80%	18.51	72.00%
	Qwen-VL-Max	27.93	28.14	28.08	14.80%	28.29	99.20%	27.68	67.80%	28.96	72.00%
	Qwen-VL-2	27.69	30.72	27.75	14.80%	30.87	99.20%	28.96	67.80%	31.13	72.00%
	GPT4-o	31.12	27.02	30.88	14.80%	26.87	99.20%	27.72	67.80%	29.10	72.00%
Dyn-VQA ch	Ds.-VL-Chat	20.74	46.91	22.37	6.38%	46.05	95.66%	44.13	84.26%	33.60	46.95%
	Qwen-VL-Max	31.53	46.73	33.53	6.38%	46.38	95.66%	44.82	84.26%	39.79	46.95%
	Qwen-VL-2	31.52	46.70	33.52	6.38%	46.28	95.66%	44.69	84.26%	39.85	46.95%
	GPT4-o	36.46	51.27	37.32	6.38%	50.85	95.66%	49.4	84.26%	42.45	46.95%
Dyn-VQA en	Ds.-VL-Chat	24.37	43.28	26.80	14.13%	42.08	89.79%	40.61	76.08%	31.67	29.51%
	Qwen-VL-Max	37.54	45.27	38.03	14.13%	44.30	89.79%	43.55	76.08%	39.40	29.51%
	Qwen-VL-2	37.37	45.16	37.25	14.13%	43.84	89.79%	43.48	76.08%	40.66	29.51%
	GPT4-o	43.33	49.71	42.40	14.13%	48.48	89.79%	47.66	76.08%	45.07	29.51%
NoCaps	Ds.-VL-Chat	43.89	40.45	43.89	0.00%	42.76	38.40%	43.89	0.00%	43.89	0.00%
	Qwen-VL-Max	37.47	34.55	37.47	0.00%	36.75	38.40%	37.47	0.00%	37.47	0.00%
	Qwen-VL-2	37.26	34.61	37.26	0.00%	36.35	38.40%	37.26	0.00%	37.26	0.00%
	GPT4-o	32.12	36.25	32.12	0.00%	33.22	38.40%	32.12	0.00%	32.12	0.00%
Visual7W	Ds.-VL-Chat	46.55	46.62	46.29	31.36%	46.03	35.37%	46.58	2.96%	46.55	0.52%
	Qwen-VL-Max	46.07	44.44	48.63	31.36%	45.16	35.37%	46.13	2.96%	46.07	0.52%
	Qwen-VL-2	45.94	43.86	48.47	31.36%	45.06	35.37%	45.99	2.96%	45.94	0.52%
	GPT4-o	41.59	37.16	40.09	31.36%	39.41	35.37%	41.80	2.96%	41.48	0.52%
Mix	Ds.-VL-Chat	25.81	35.23	26.55	12.67%	**35.38**	76.83%	33.06	49.33%	32.73	38.33%
	Qwen-VL-Max	32.35	34.78	33.00	12.67%	35.48	76.83%	34.84	49.33%	35.51	38.33%
	Qwen-VL-2	32.59	35.56	33.27	12.67%	36.29	76.83%	35.62	49.33%	36.33	38.33%
	GPT4-o	34.52	35.96	33.99	12.67%	35.90	76.83%	35.86	49.33%	36.49	38.33%

Table 7: Knowledge Boundary model as a surrogate boundary identifier for other VLLMs. Results evaluated by token accuracy.

		No RAG	All RAG	Human	%	HKB	%	SKB	%
MMMU	Qwen-VL-Chat	20.12	20.28	**21.24**	6.88%	20.35	97.08%	20.18	61.26%
	Qwen-VL-Max	51.33	41.37	**52.67**	6.88%	41.46	97.08%	44.40	61.26%
	Qwen-VL-2	51.45	42.39	**51.93**	6.88%	42.54	97.08%	45.61	61.26%
	GPT4-o	56.60	56.64	**57.36**	6.88%	56.92	97.08%	56.91	61.26%

Table 8: Results evaluated by LLM on MMMU validation set.

present the performance of each of the 30 subjects in the MMMU validation set in Fig 4. The first row shows the LLM evaluation results and the second shows the token accuracy metric. We can see that in most subjects “Human” setting succeeds in obtaining a higher performance than both “All RAG” and “No RAG” settings.

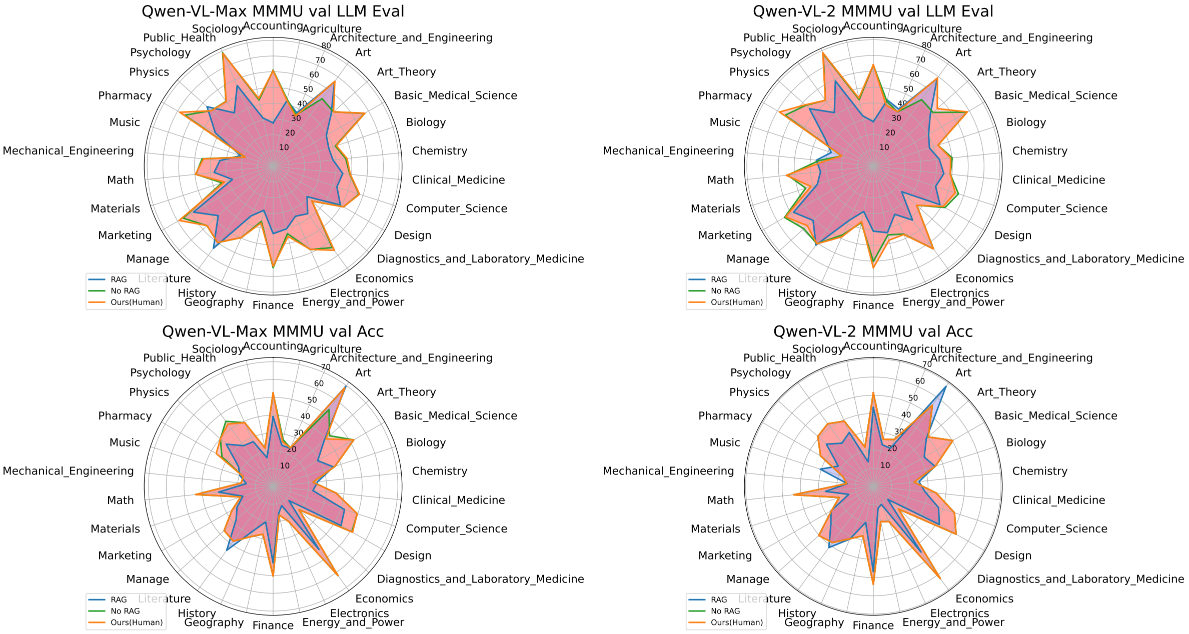

Figure 4: Qwen-VL-Max and Qwen-VL-2 performance on MMMU validation set with Knowledge Boundary model trained on Human-labeled data.