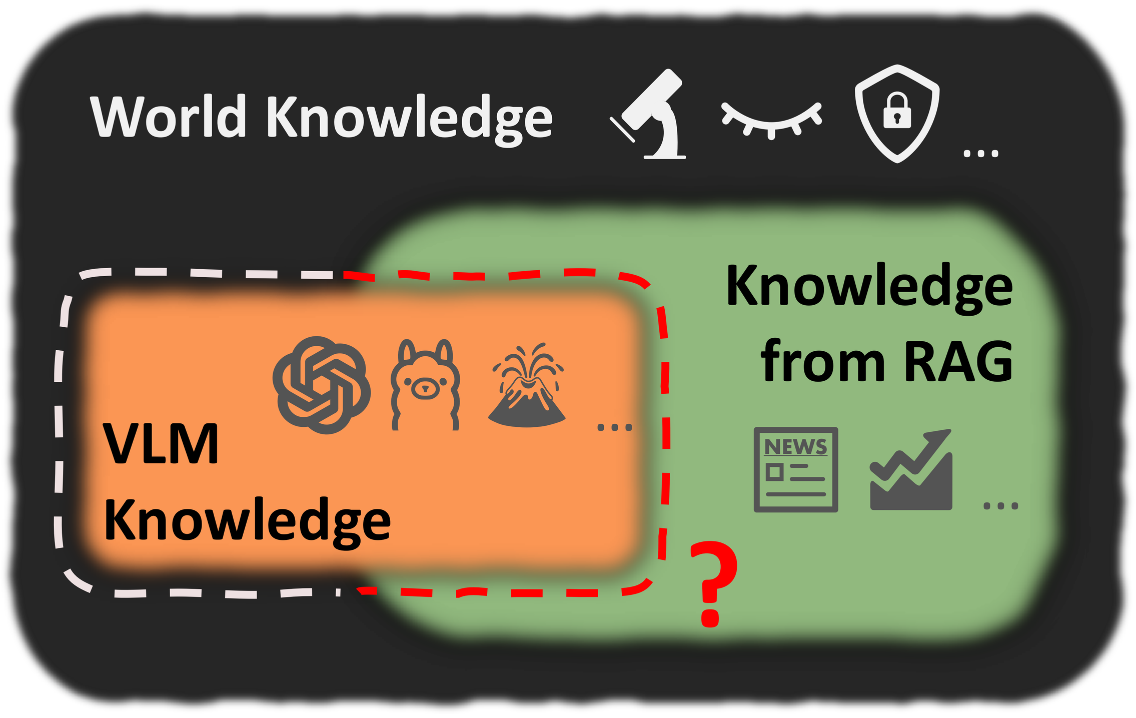
Detecting Knowledge Boundary of Vision Large Language Models by Sampling-Based Inference

上海科技大学 ShanghaiTech University × 通义

Zhuo Chen, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinyu Geng, Pengjun Xie, Fei Huang, Kewei Tu

ShanghaiTech University & Institute for Intelligent Computing Alibaba Group

Overview



Aim to identify the VLM knowledge boundary.

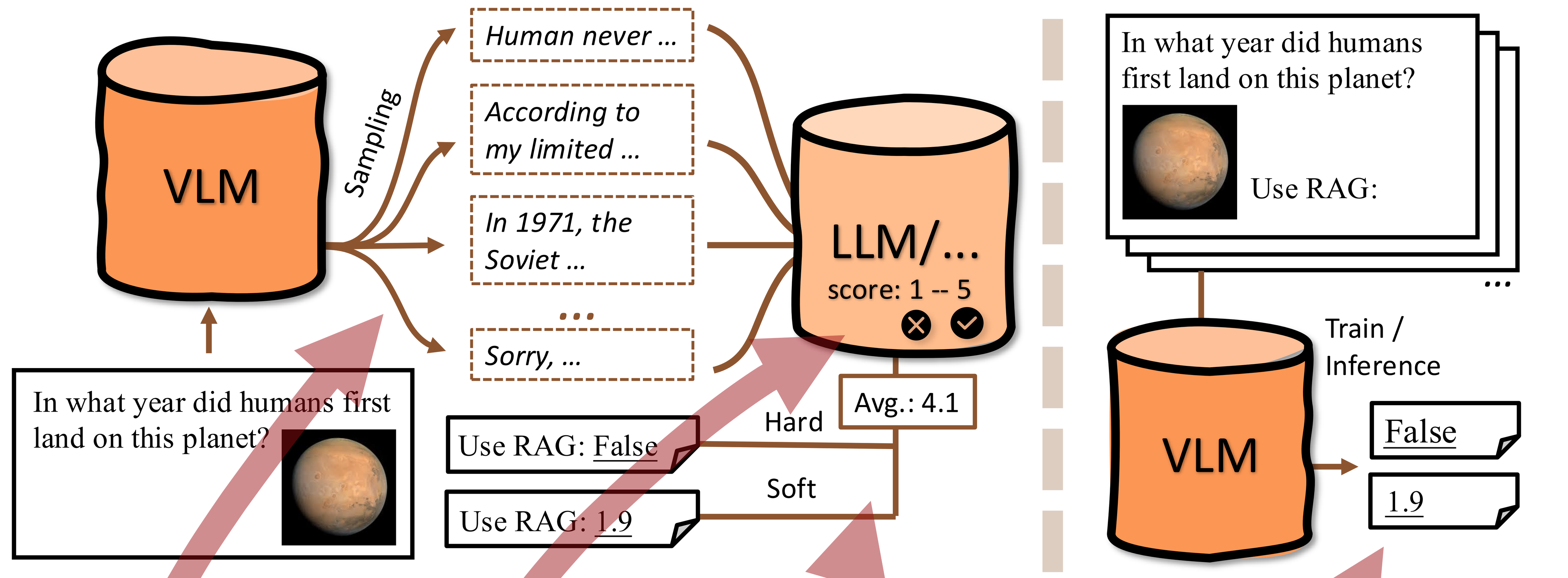
Method with two variants that fine-tune a VLM on an automatically constructed dataset for boundary identification.

Experiments validate the potential VQA performance improvements using RAG.

Code

Paper

Method



Sample a VLM and collect its output

Evaluate the correctness by a text LLM

Label the query with "Using RAG" or not with two variants

1. a soft score showing tendency
2. a hard True/False label

Train a VLM to predict the **label**

At inference time, we can selectively use RAG conditioned on the prediction thus obtain higher efficiency.
(Soft version: by setting a threshold)

Experiment Setup

Model to be sampled (30 times) and fine-tuned: **Qwen-VL-7B-Chat**.
First, evaluate the **original model's** performance that are sampled.
Second, validate whether the identified knowledge boundary can function as a **surrogate boundary** for other VLMs
Training and test data are completely **non-overlapping**. Five test datasets are **diverse.**

Source	# Samples	Model	Avg. Score ± std.
InfoSeek	216000	QW DS	1.82± 1.17 1.86± 1.28
OK-VQA	9009	QW DS	3.70± 1.48 4.92± 0.47
VQAv2.0	108000	QW DS	4.27± 1.36 4.50± 1.22
MMBench (en)	4329	QW DS	3.92± 1.72 4.08± 1.65
MME	2374	QW DS	4.15± 1.63 4.15± 1.64

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

Experiments -Main Result

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%

% shows the portion of each dataset that adopts RAG. (No RAG=0%; All RAG=100%)
Ideally, we want to **maximize performance** and **minimize RAG%.**

Knowledge Boundary as a Surrogate

	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%
	GPT-4o	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%
	GPT-4o	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%
	GPT-4o	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%
	GPT-4o	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%
	GPT-4o	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%
	GPT-4o	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%
	GPT-4o	51.44	52.90	50.57	12.67%	54.10	76.83%	52.97	49.33%	55.27	38.33%

Case Study

Question: How many World Series titles has the team won?

Prediction w/o retrieval: The Houston Astros have won one World Series title, in 2017. → (incorrect)
Prediction w/ retrieval: The Houston Astros have won two World Series titles. → (correct) *(Source: Dyn-VQA (en))*

Question: Who is the current Prime Minister of this country?

Prediction w/o retrieval: The current Prime Minister of the United Kingdom is Rishi Sunak. → (incorrect)
Prediction w/ retrieval: The current Prime Minister of the United Kingdom is Keir Starmer. → (correct) *(Source: Dyn-VQA (en))*

Cases where model predicts to be **out of** the knowledge boundary.
→ RAG helps **correct** the answer.

Question: Can you add a caption to the image using a phrase? For example: A little girl in a white jacket and sandals.

Prediction w/o retrieval: A blue jay perched on a tree branch. → (correct)
Prediction w/ retrieval: A blue and white bird perched on a tree branch. → (correct) *(Source: NoCaps)*

Question: How is the boat staying ashore?

Prediction w/o retrieval: The boat is staying ashore by being tied to a wooden post with a yellow rope. The rope is wrapped around the post and secured, preventing the boat from drifting back into the water. → (correct)
Prediction w/ retrieval: It is tied to a tree with rope. → (correct) *(Source: Visual7W)*

Cases where the model predicts to be **in** the knowledge boundary
→ We can **save the time** spent on RAG.