Self-Knowledge Guided Retrieval Augmentation for Large Language models

Name
박제현
NLP
2024/05/21

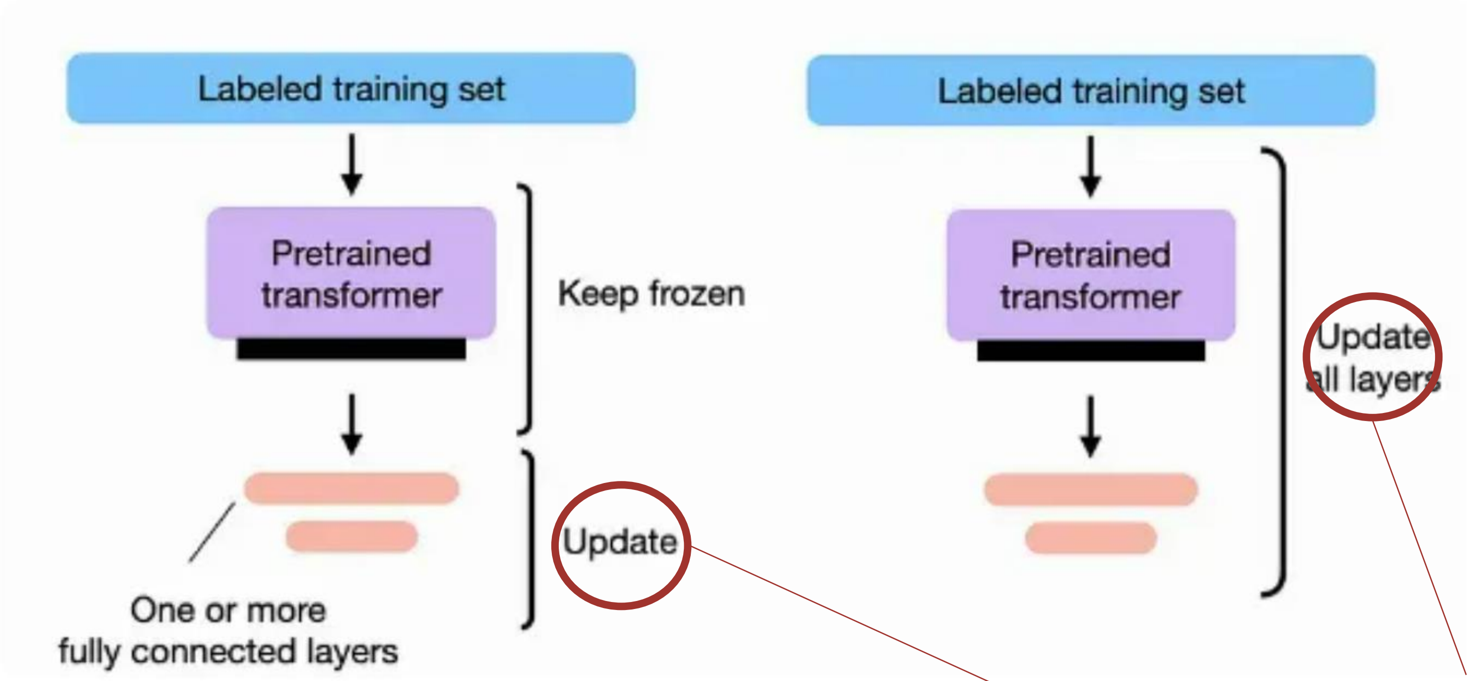
TRAIN AND TEST

Contents

TRAIN AND TEST

Introduction - LLMs & Retrieval Knowledge

Fine-tuning

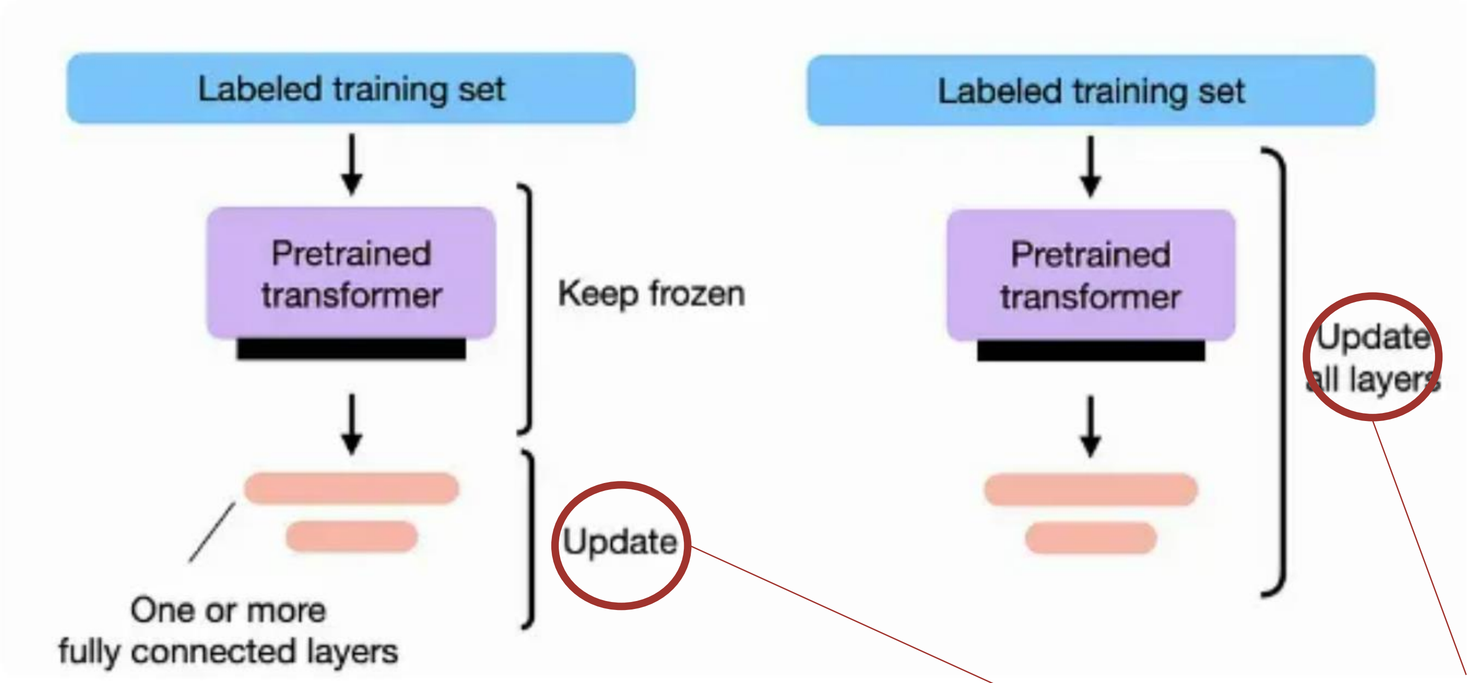


Not efficient for specific-domain
Fine-tuning for specific domains

TRAIN AND TEST

Introduction - LLMs & Retrieval Knowledge

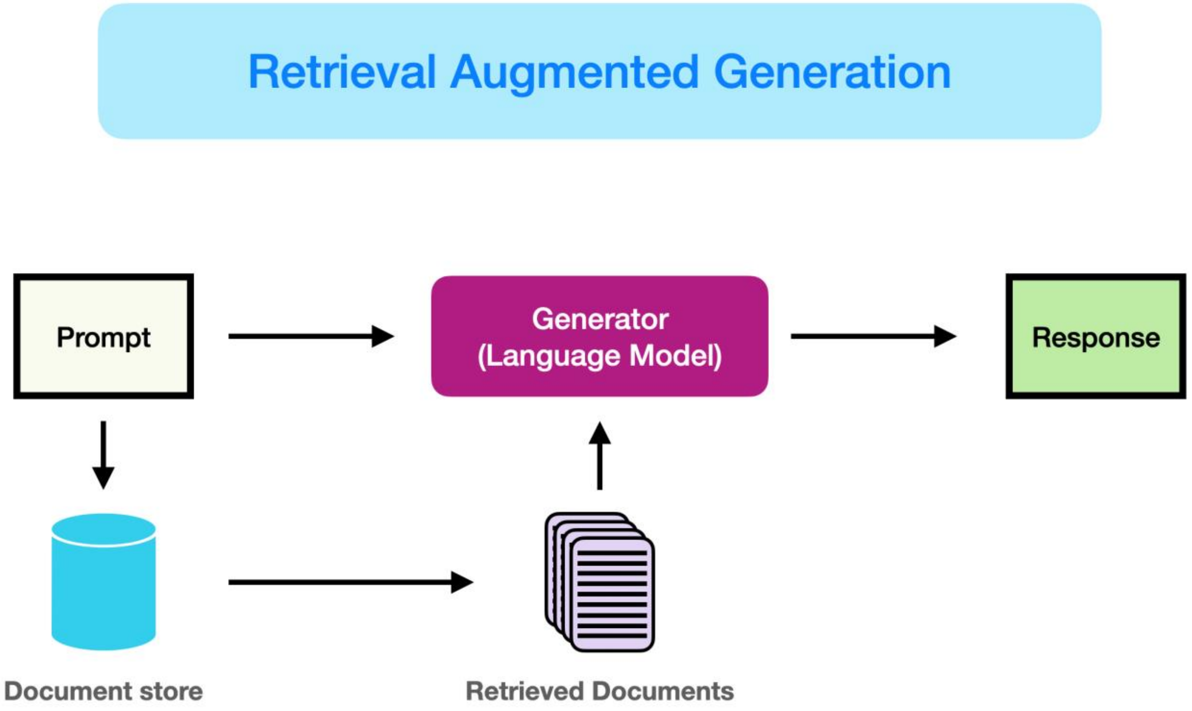
Fine-tuning



Difficult to update for
time&computational costs

Introduction - LLMs & Retrieval Knowledge

RAG



Using Query for better use of LLMs
Find the **meaningful vector**

Introduction - LLMs & Retrieval Knowledge

RAG-advance

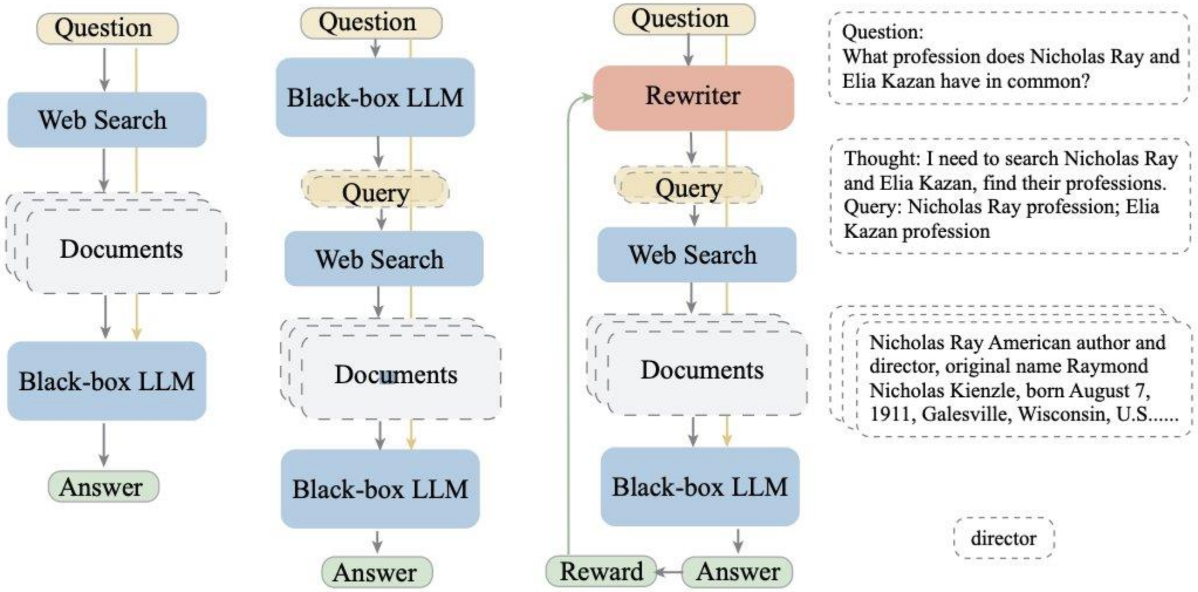


Figure 1: Overview of our proposed pipeline. From left to right, we show standard *retrieve-then-read* method, LLM as a query rewriter and *rewrite-retrieve-read* pipeline with a trainable rewriter.

Using Query for better use of LLMs

- **Prompt-Engineering** / **Rewards**

TRAIN AND TEST

Introduction - LLMs & Retrieval Knowledge

Limitations

LLMs are becoming **more knowledgeable**
Original method **does not improve**

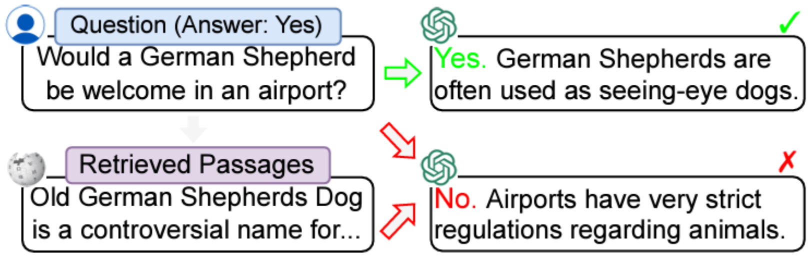


Figure 1: Comparison between two responses given by InstructGPT. The retrieved passages are relevant but not particularly helpful for solving the question, which influences the model's judgment and leads to incorrect answers.

it is **distracted** and gives incorrect ones by ~~**adding retrieved passages**~~.

TRAIN AND TEST

Introduction - LLMs & Retrieval Knowledge

What is the Problem?

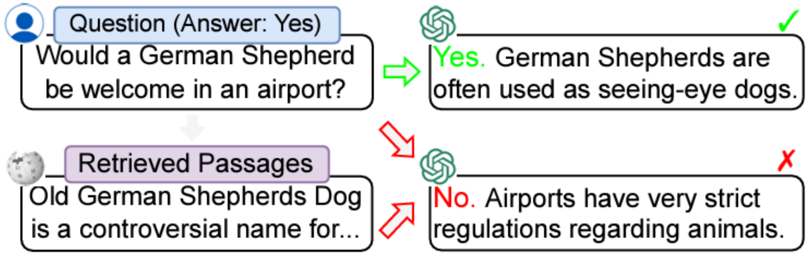
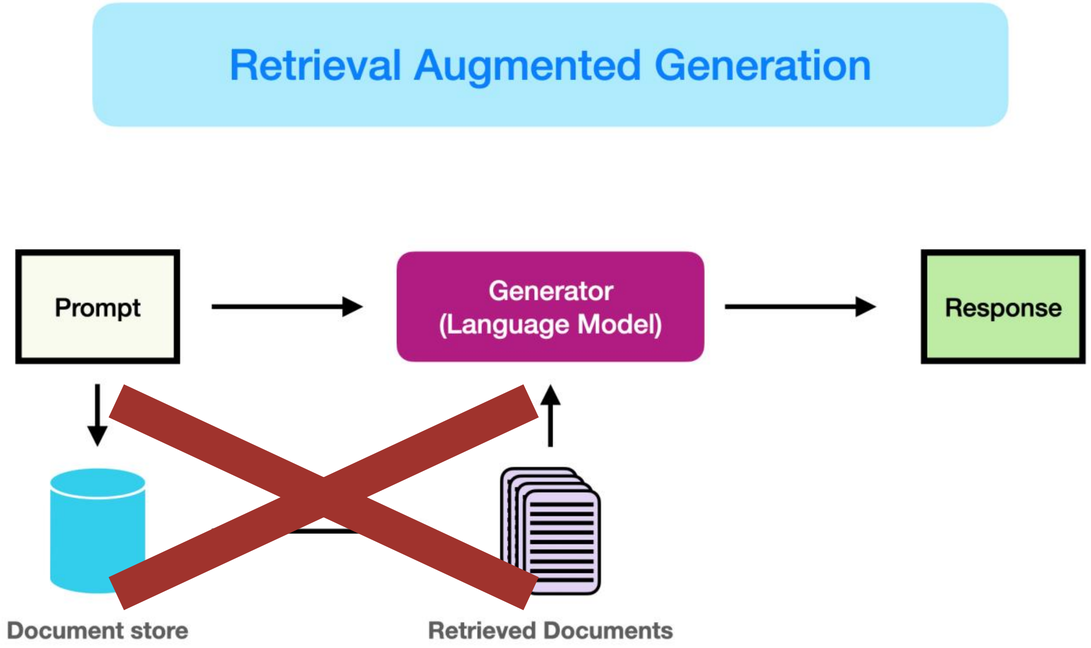


Figure 1: Comparison between two responses given by InstructGPT. The retrieved passages are relevant but not particularly helpful for solving the question, which influences the model's judgment and leads to incorrect answers.

it is **distracted** and gives incorrect ones by ~~adding retrieved passages~~.



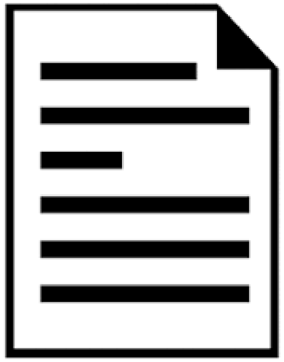
It is difficult to know **in advance** whether the **retrieved results** are better than **what LLMs already captured**.

TRAIN AND TEST

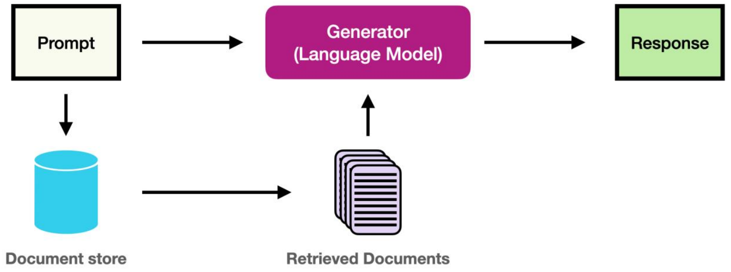
Introduction - LLMs & Retrieval Knowledge

Ideation

Study documents for what I don’t know



Retrieval Augmented Generation



What is Self-Knowledge : Language Models (Mostly) Know What They Know

Self-Knowledge

Glossary: Observables and Metrics

• **P(True)** – The probability a model assigns to the **proposition that a specific sample is the correct answer** to a question.

• **P(IK)** – **The probability a model assigns to "I know"**, i.e. the proposition that **it will answer a given question correctly** when samples are generated at unit temperature. In this work, P(IK) is usually computed using a **binary classification head on top of a language model**.

• **Ground Truth P(IK)**– The fraction of unit temperature samples to a question that are correct.

TRAIN AND TEST

What is Self-Knowledge : Language Models (Mostly) Know What They Know

Self-Knowledge

• **P(IK)** – **The probability a model assigns to "I know",** the proposition **it will answer a given question correctly**

P(IK) is usually computed using a
binary classification head on top of a language model.

Tokens, BCE, Embedding space, etc . . .

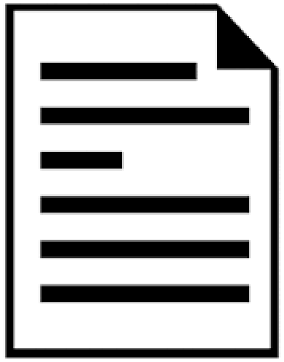
Figure 3 Examples of P(IK) scores from a 52B model. Token sequences that ask harder questions have lower P(IK) scores on the last token. To evaluate P(IK) on a specific full sequence, we simply take the P(IK) score at the last token. Note that we only train P(IK) on final tokens (and not on partial questions).

TRAIN AND TEST

Introduction - LLMs & Retrieval Knowledge

SKR

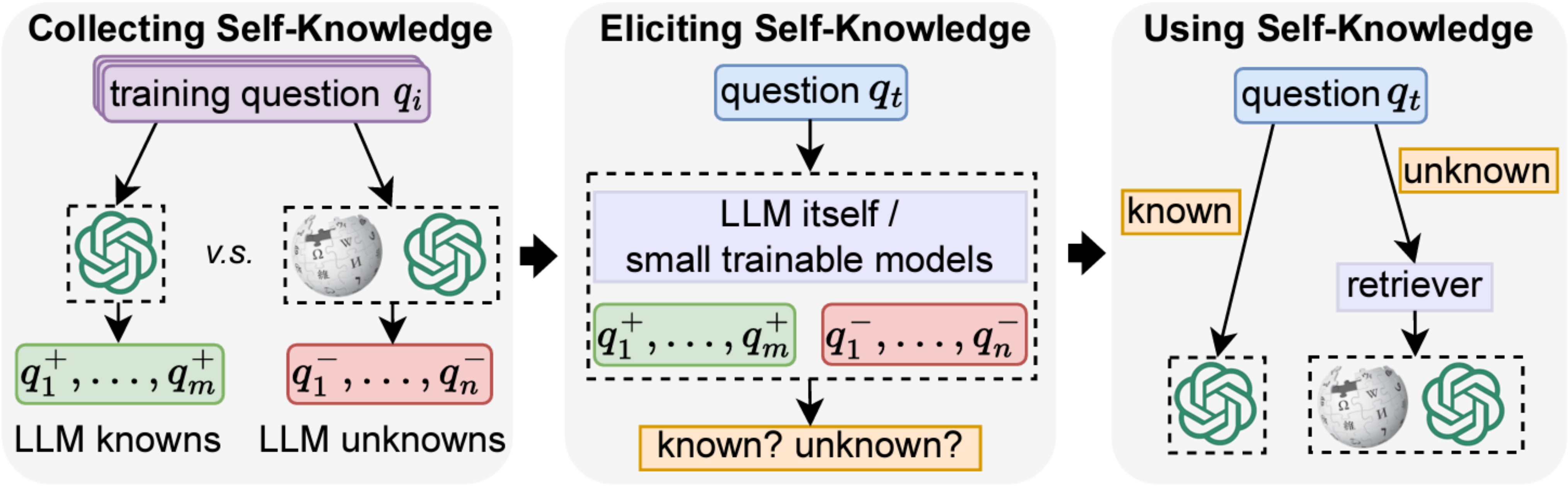
introducing the **external resources**, we **detect the knowledge boundary** of LLMs through the performance changes.



Experimental results show that SKR outperforms **chain-of-thought based** (Wei et al., 2022) and **fully retrieval-based methods** by 4.08%/2.91% (for InstructGPT) and 4.02%/4.20% (for ChatGPT), respectively

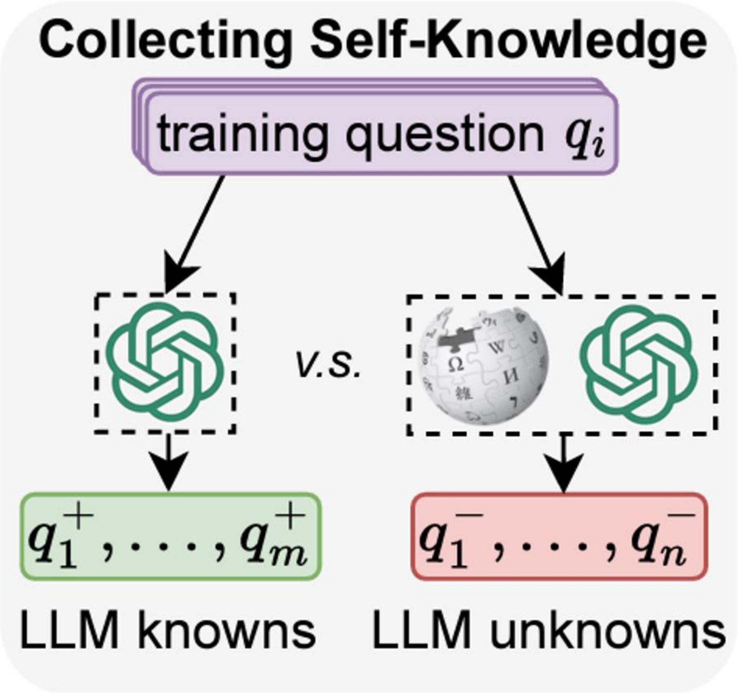
Approach - Pipeline

Pipeline



Approach – Collecting Self-Knowledge

Collecting Self-Knowledge



Reflects the internal knowledge to question qi in M

$$\hat{a}(\mathcal{M}, q_i) = \mathcal{M}(q_1 \circ a_1, \ldots, q_d \circ a_d, q_i),$$

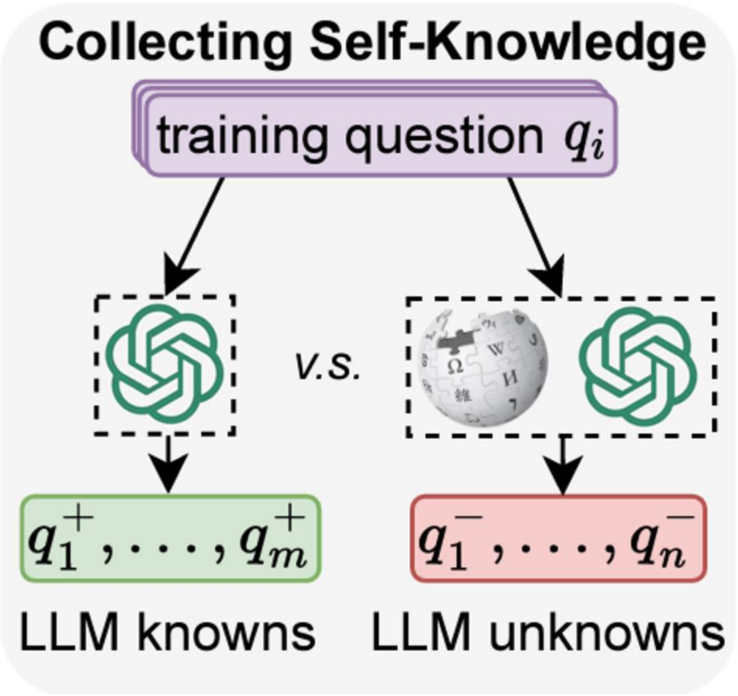
Given a dataset D with training **question-answer pairs {qj,aj}|D| j=1**, we can use the LLM M to generate the answers for each question qi via **few-shot in-context learning**

$$\{q_j \circ a_j\}_{j=1}^{d}$$

TRAIN AND TEST

Approach – Collecting Self-Knowledge

Collecting Self-Knowledge



Dense Passage Retrieval for Open-Domain Question Answering

$$p_i = \{p_{i1}, p_{i2}, \ldots, p_{ik}\} = \mathcal{R}(q_i, \mathcal{C}),$$

Top-k retrieved passages for qi

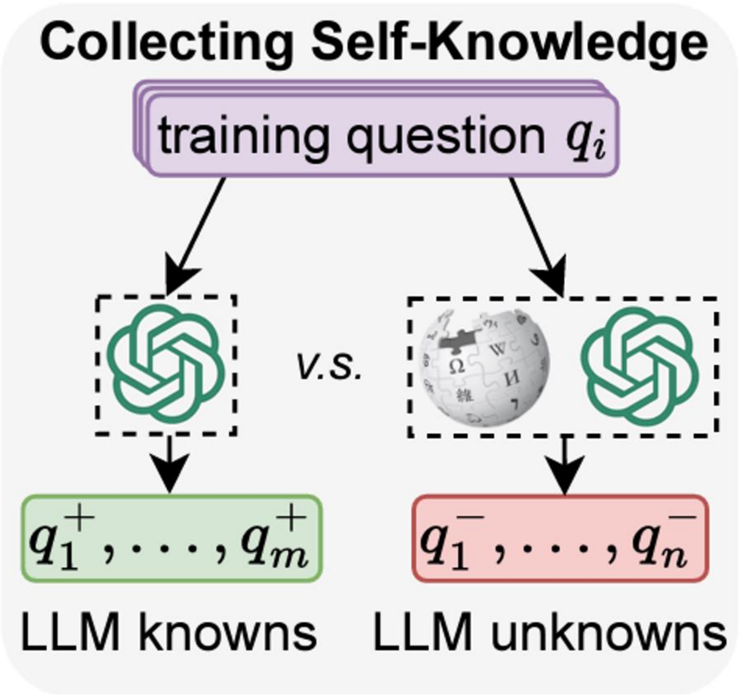
$$\hat{a}^{\mathcal{R}}(\mathcal{M}, q_i) = \mathcal{M}(q_1 \circ p_1 \circ a_1, \ldots, q_d \circ p_d \circ a_d, q_i \circ p_i).$$

$$\hat{a}(\mathcal{M}, q_i),\ \hat{a}^{\mathcal{R}}(\mathcal{M}, q_i),$$

TRAIN AND TEST

Approach – Collecting Self-Knowledge

Collecting Self-Knowledge



$$q_i \in \begin{cases} \mathcal{D}^+, & \text{if } \mathrm{E}[\hat{a}(\mathcal{M}, q_i)] \geq \mathrm{E}[\hat{a}^{\mathcal{R}}(\mathcal{M}, q_i)]; \\ \mathcal{D}^-, & \text{otherwise}, \end{cases}$$

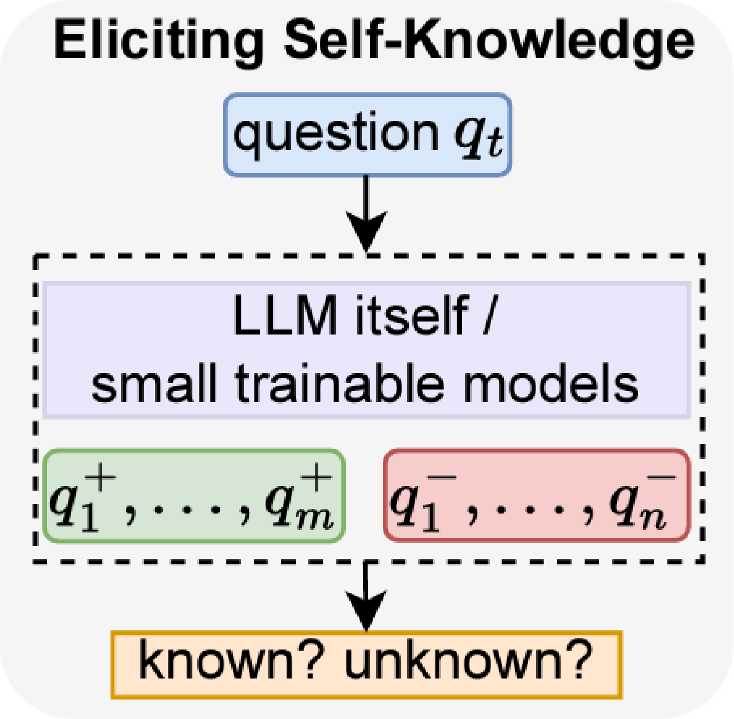
E is an evaluation metric such as accuracy and exact match score

What is the Evaluation Metric? <- Key but couldn't find

TRAIN AND TEST

Approach – Eliciting Self-Knowledge of LLMs

Eliciting Self-Knowledge of LLMs



Four different strategies are proposed to detect the self-knowledge of target questions, including **direct prompting**, **in-context learning**, **training a classifier**, and **nearest neighbor search**

Direct Prompting

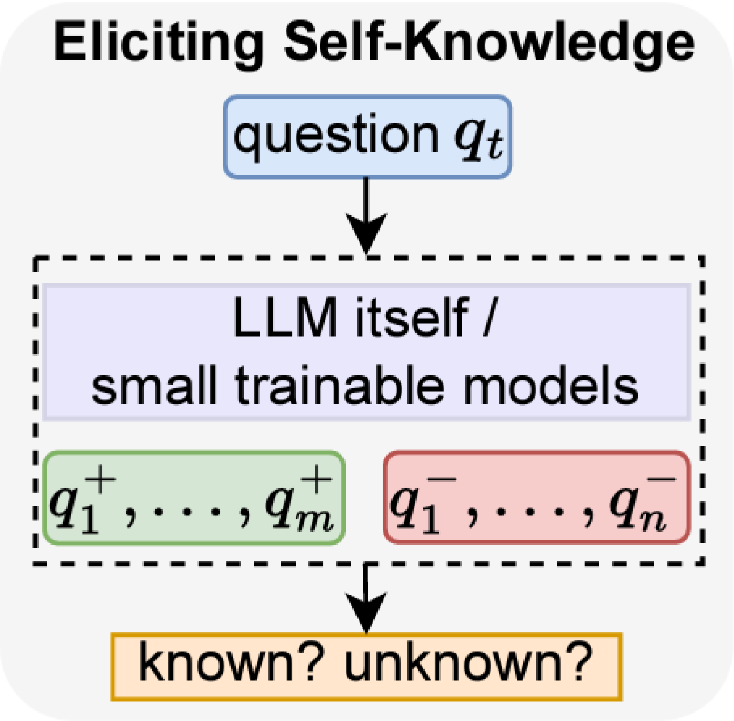
(prompt)
$\{q_t\}$ Q: *Do you need additional information to answer this question?* A:

(possible response)
No, I don't need additional information to answer this question. / Yes, I need additional information to answer this question.

Tests each question independently and does not make use of the collected training questions

Approach – Eliciting Self-Knowledge of LLMs

Eliciting Self-Knowledge of LLMs



Four different strategies are proposed to detect the self-knowledge of target questions, including **direct prompting**, **in-context learning**, **training a classifier**, and **nearest neighbor search**

In-Context Learning

(prompt)

$\{q_1^+\}$ Q: *Do you need additional information to answer this question?* A: *No, I don't need additional information to answer this question.*

$\{q_1^-\}$ Q: *Do you need additional information to answer this question?* A: *Yes, I need additional information to answer this question.*

......

$\{q_t\}$ Q: *Do you need additional information to answer this question?* A:

(possible response)

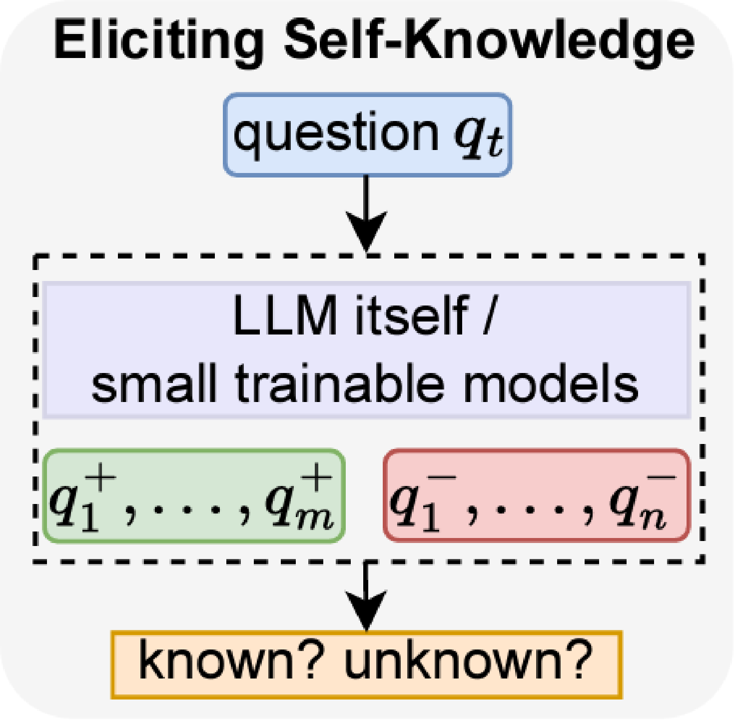
No, I don't need additional information to answer this question. / Yes, I need additional information to answer this question.

Both methods require designing prompts and calling the LLMs for each new question, which makes it impractical. Unstable due to contextual bias and sensitivity for close-source LLMs

TRAIN AND TEST

Approach – Eliciting Self-Knowledge of LLMs

Eliciting Self-Knowledge of LLMs



Four different strategies are proposed to detect the self-knowledge of target questions, including **direct prompting**, **in-context learning**, **training a classifier**, and **nearest neighbor search**

Two-way classification problem

$$q_i \in \mathcal{D}^+ \cup \mathcal{D}^-$$

$$\hat{y}_i = \text{softmax}(W h_{\text{cls}}(q_i) + b),$$

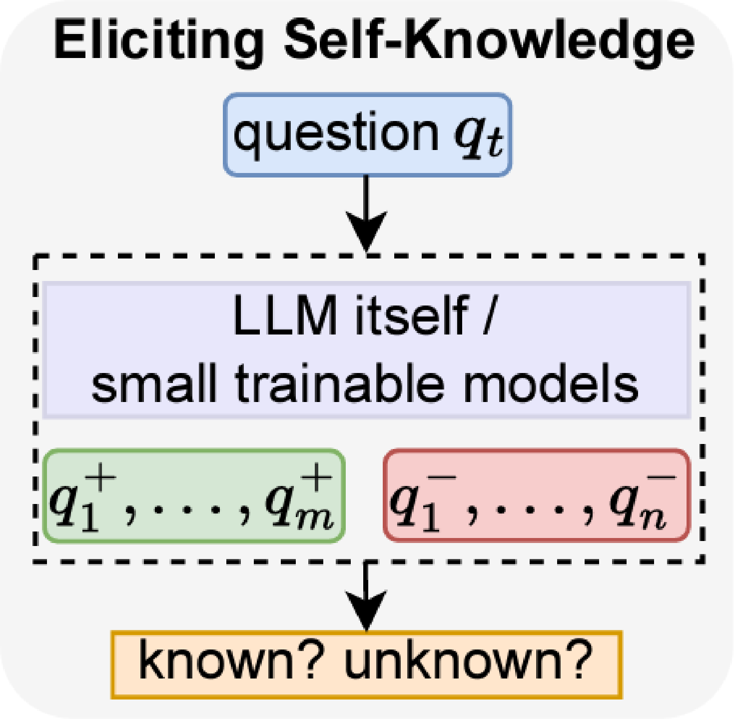
BERT-base

Minimizing the **cross-entropy loss** between
predicted label distribution ˆyi and **ground-truth label of qi**

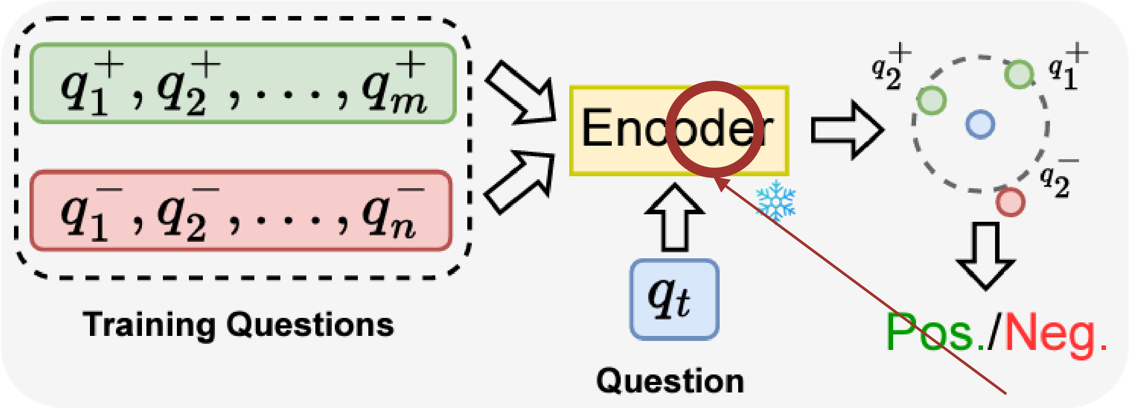
TRAIN AND TEST

Approach – Eliciting Self-Knowledge of LLMs

Eliciting Self-Knowledge of LLMs



Four different strategies are proposed to detect the self-knowledge of target questions, including **direct prompting**, **in-context learning**, **training a classifier**, and **nearest neighbor search**



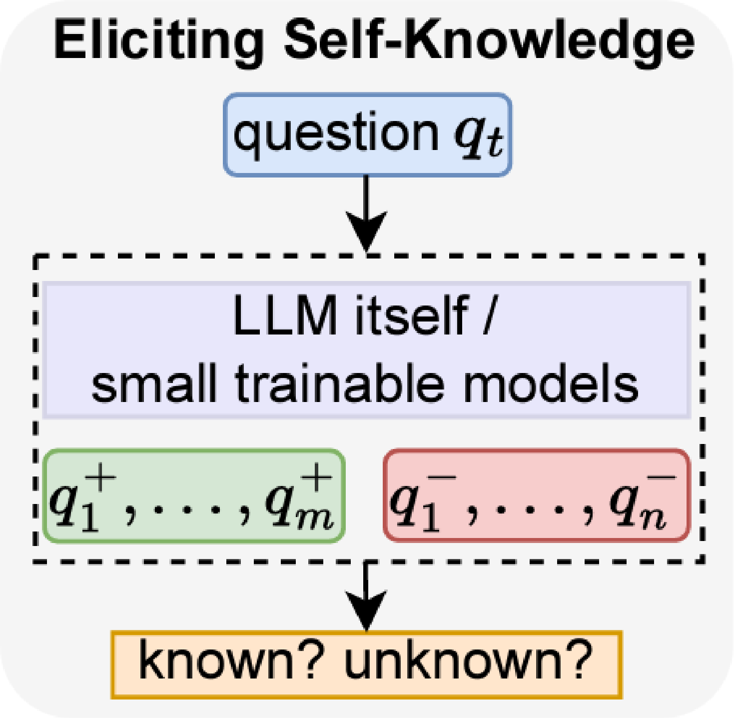
pre-trained fixed encoder

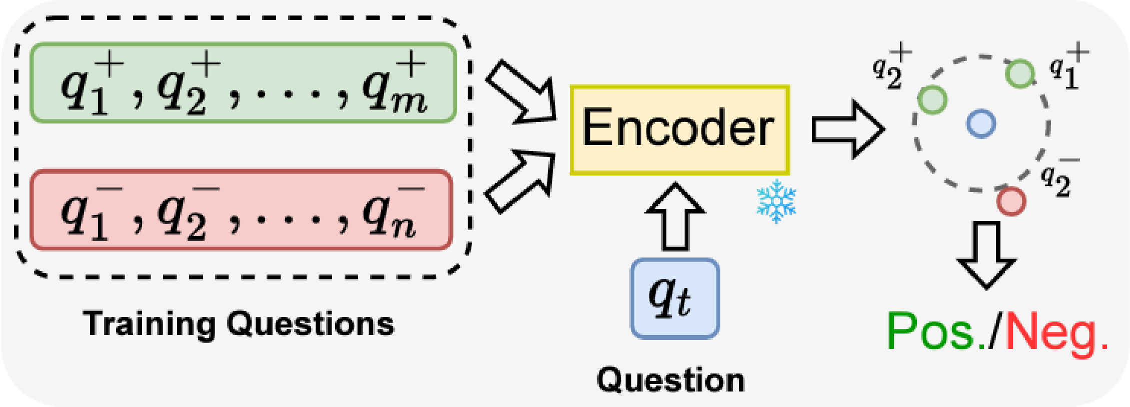
$$\mathrm{sim}(q_t, q_i) = \frac{e(q_t) \cdot e(q_i)}{||e(q_t)|| \cdot ||e(q_i)||}$$

TRAIN AND TEST

Approach – Eliciting Self-Knowledge of LLMs

Eliciting Self-Knowledge of LLMs

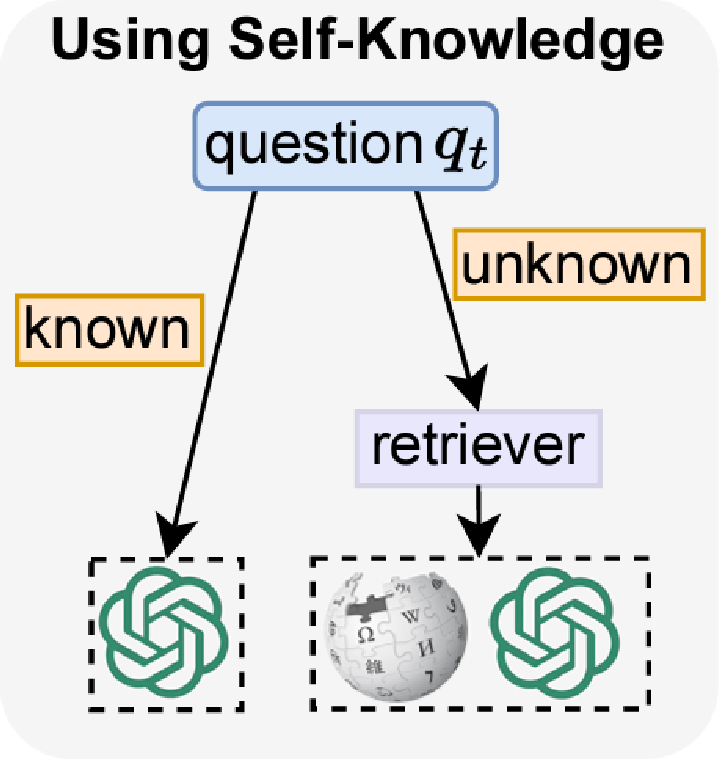




Top-k nearest neighbors include **ℓ positive** ones and **k −ℓ negative** ones,
positive if ℓ / k−ℓ ≥ m / n
negative if ℓ / k−ℓ < m / n
(**m** and **n** are the numbers of questions in **D+** and **D−**respectively).

Approach – Using Self-Knowledge for Adaptive Retrieval Augmentation

Using Self-Knowledge



Adaptive Retrieval Augmentation

(for LLM known)
$\{q_1 \circ a_1\}, \ldots, \{q_d \circ a_d\}, \{q_t\}$
A: (LLM directly answers without retrieval)

(for LLM unknown)
$\{q_1 \circ p_1 \circ a_1\}, \ldots, \{q_d \circ p_d \circ a_d\}, \{q_t\}$
Here are some passages: $\{p_t\}$
A: (LLM answers with retrieval augmentation)

TRAIN AND TEST

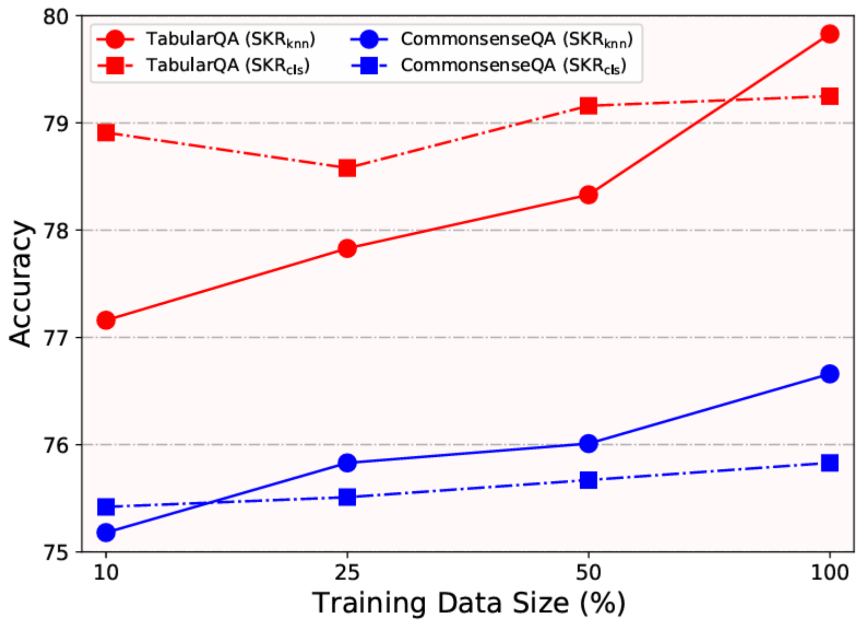
Main results & Analysis

Main results

- Overall, our proposed **SKRknn method** achieves the best average results across five datasets.

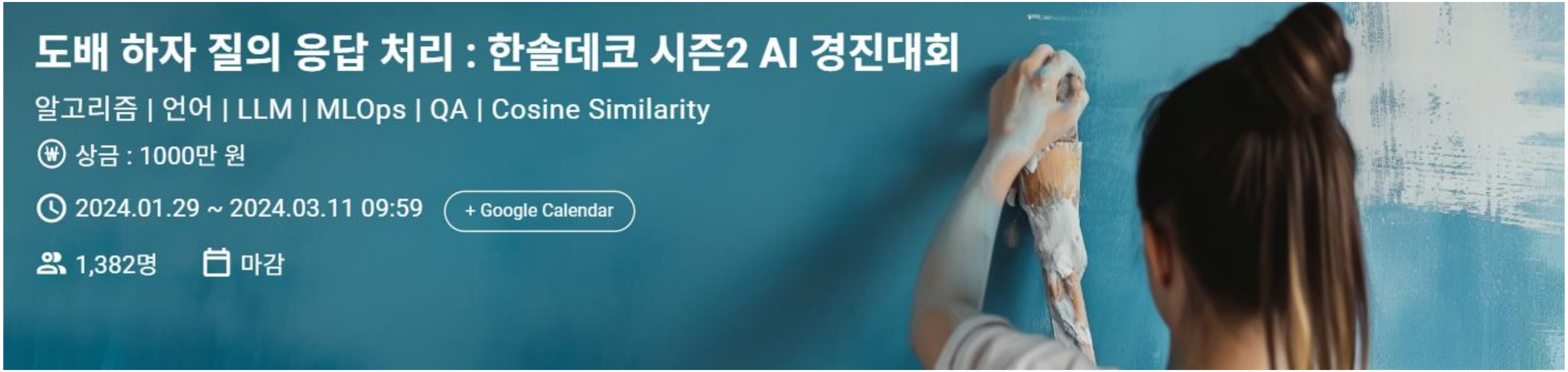
- **SKRprompt** shows relatively **poor results.**

- **SKRicl** and **SKRcls** work but **do not show consistent improvement.**



TRAIN AND TEST

Conclusion



TRAIN AND TEST

TRAIN AND TEST