

Unveiling Knowledge Boundary of Large Language Models for Trustworthy Information Access

Yang Deng¹, Moxin Li², Liang Pang³, Wenxuan Zhang⁴, Wai Lam⁵

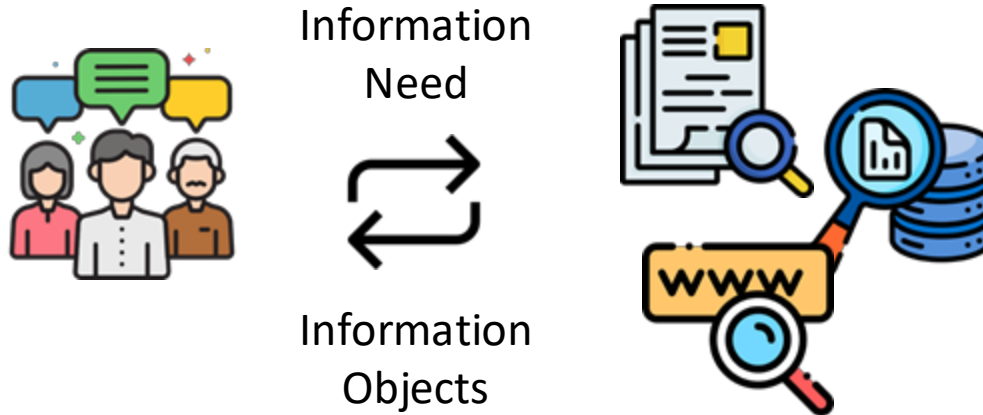
¹Singapore Management University ²National University of Singapore

³Institute of Computing Technology, Chinese Academy of Sciences

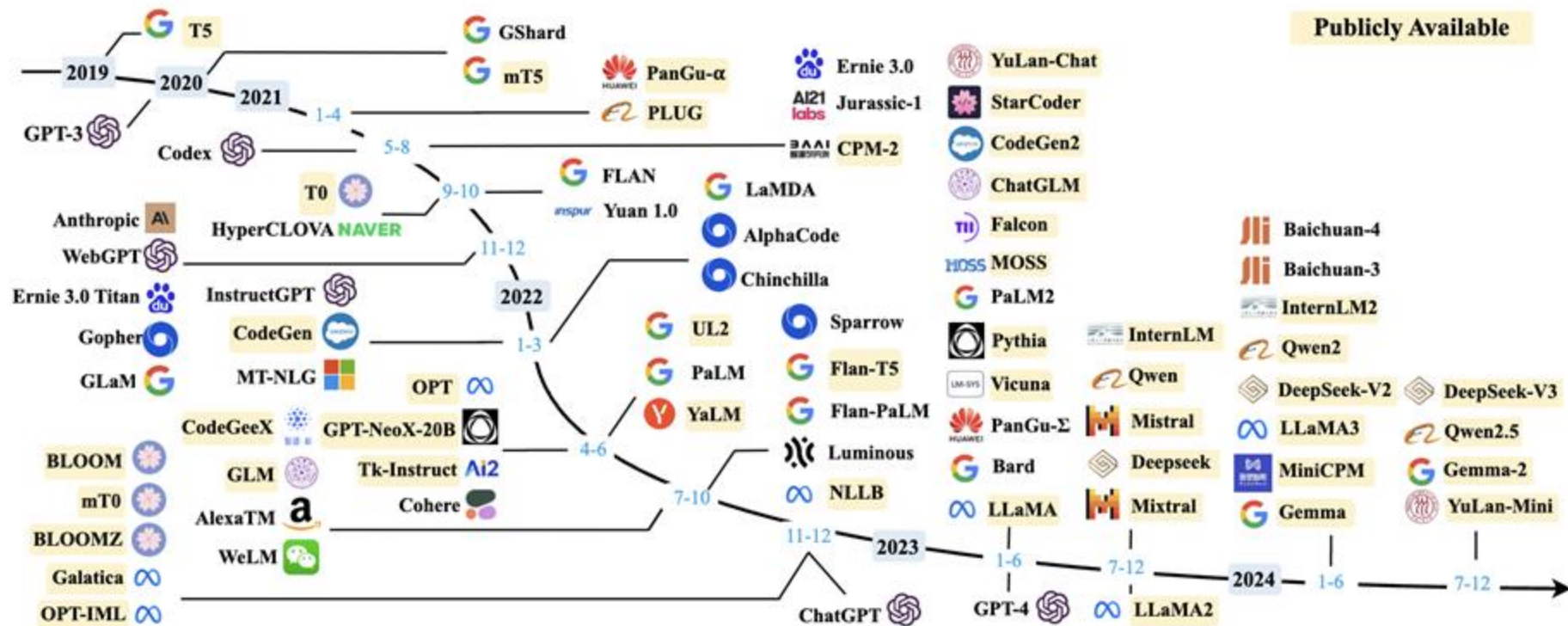
⁴Singapore University of Technology and Design ⁵The Chinese University of Hong Kong

Information Access

Information access refers to the processes and technologies that enable users to locate, retrieve, and use information from various sources (e.g., documents, databases, the web).

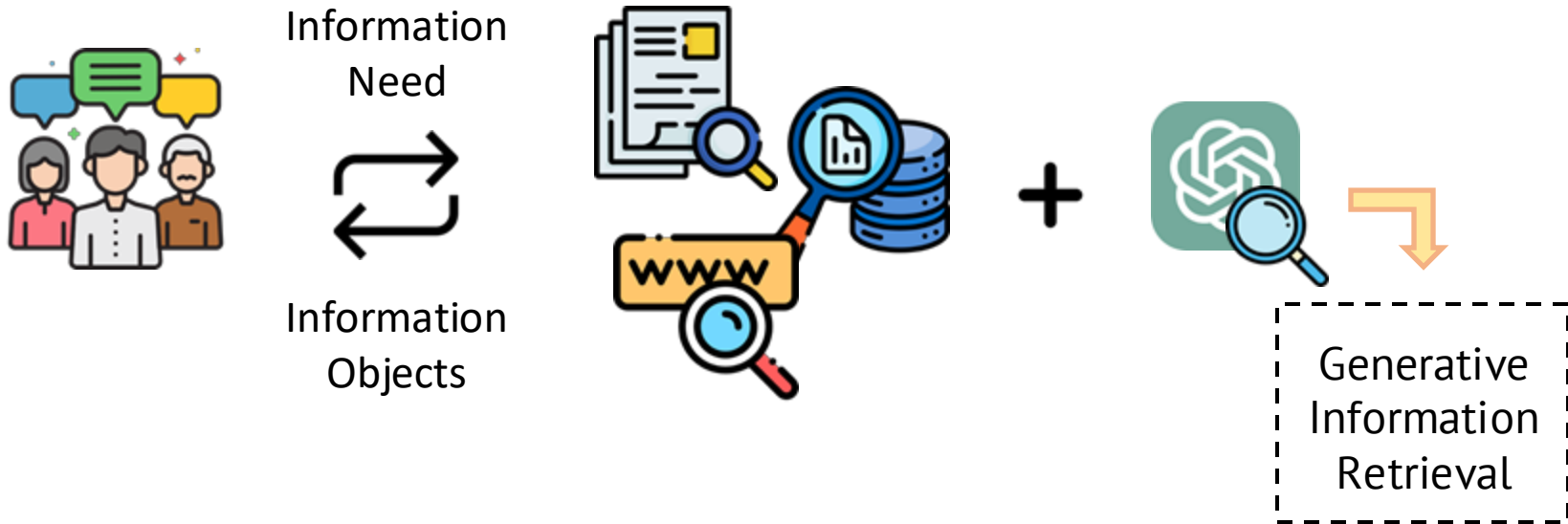


Large Language Models

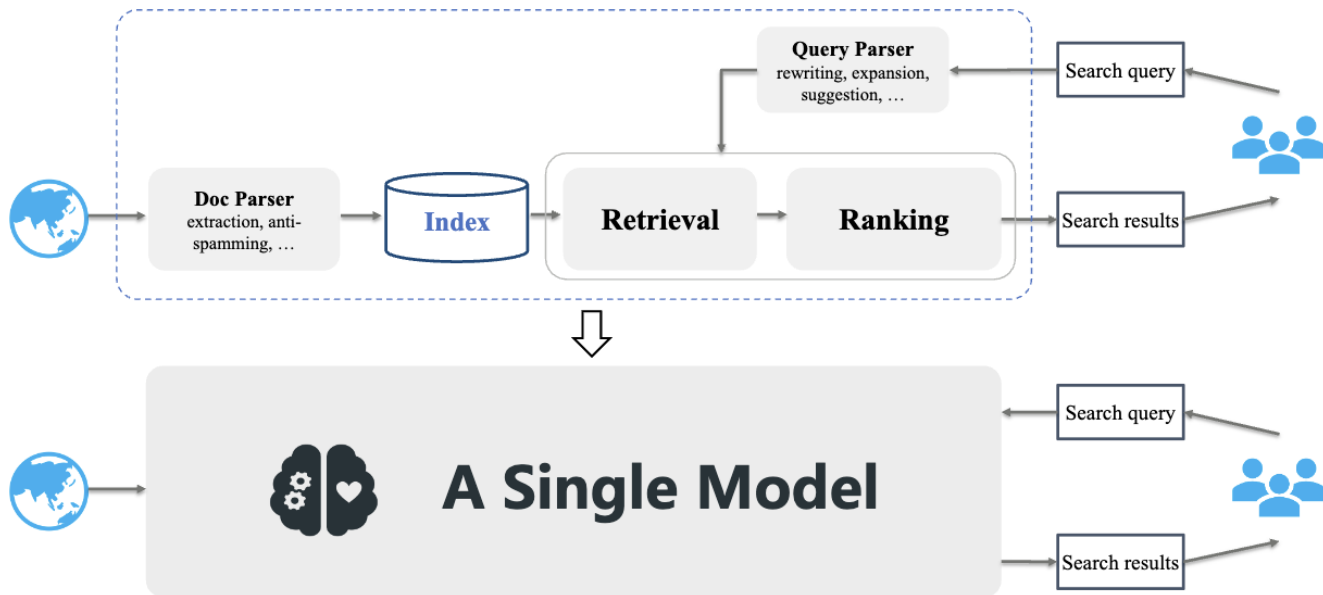


LLMs as A New Information Source

Information access refers to the processes and technologies that enable users to locate, retrieve, and use information from various sources (e.g., documents, databases, the web, **LLMs**).







The Rise of Generative Information Retrieval



Effectiveness

- ❑ Knowledge of all documents in corpus is encoded into model parameters, which can be optimized directly in an end-to-end manner
- ❑ Directly generates precise and coherent answers, rather than simply returning a list of documents, reducing the cognitive load on users.

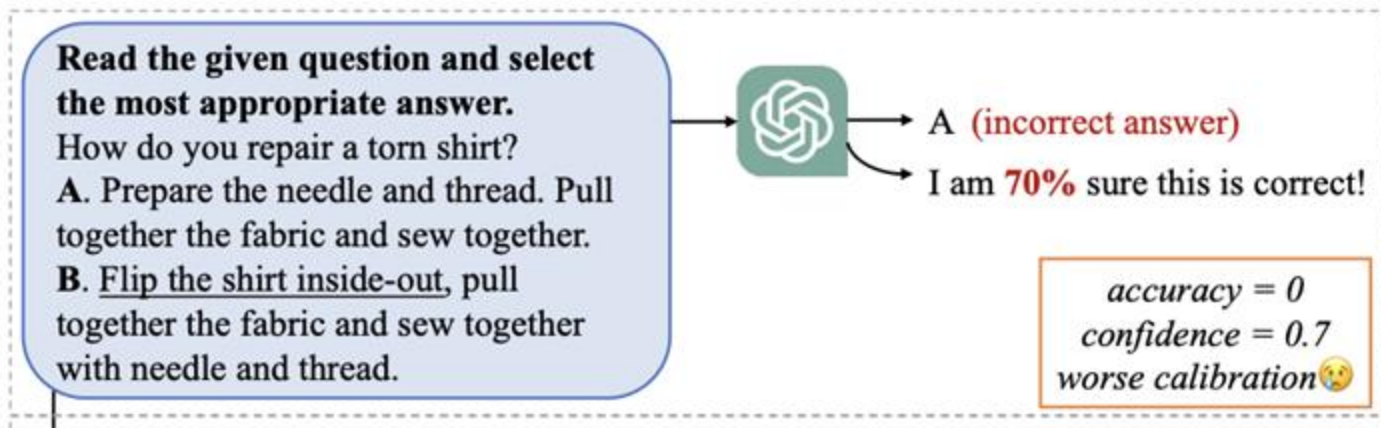
The Rise of Generative Information Retrieval

	Dense retrieval	Generative retrieval
Memory size (MS MARCO 300K)	 GTR 1430MB	 GenRet 860MB
Online latency	 GTR 1.97s	 GenRet 0.16s

Efficiency

- ❑ Generative models store knowledge within model parameters, reducing the overall memory size compared to indexing huge document collections.
- ❑ Shortens total session time, enhancing perceived latency efficiency from a user-experience perspective.

Concerns on Trustworthiness and Reliability



What's the key to a delicious pizza sauce?

Add non-toxic glue for tackiness



What's your confidence?

100%



Overconfidence



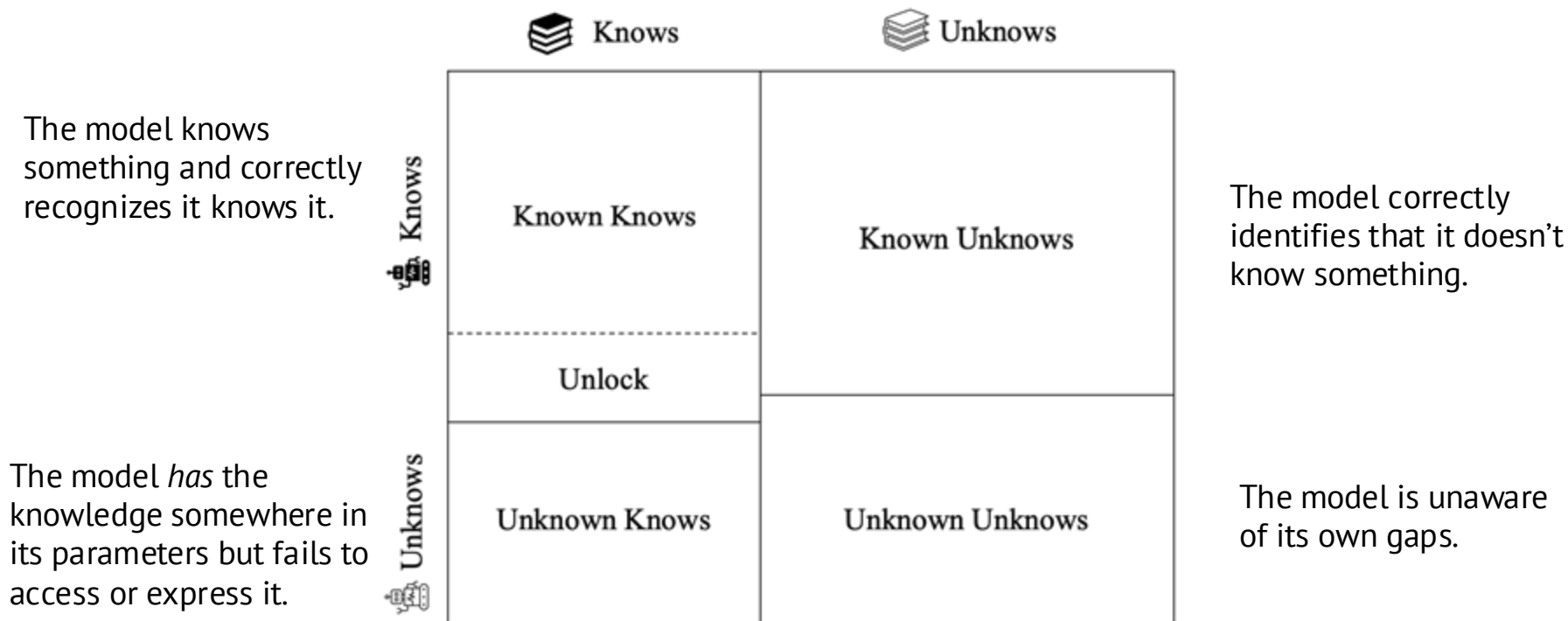
Do LLMs Know What They Don't Know?

Kapoor et al., "Large Language Models Must Be Taught to Know What They Don't Know" (NeurIPS '24)

Li et al., "Think Twice Before Assure: Confidence Estimation for Large Language Models through Reflection on Multiple Answers" (EMNLP '24 Findings)

Known-Unknown Quadrant

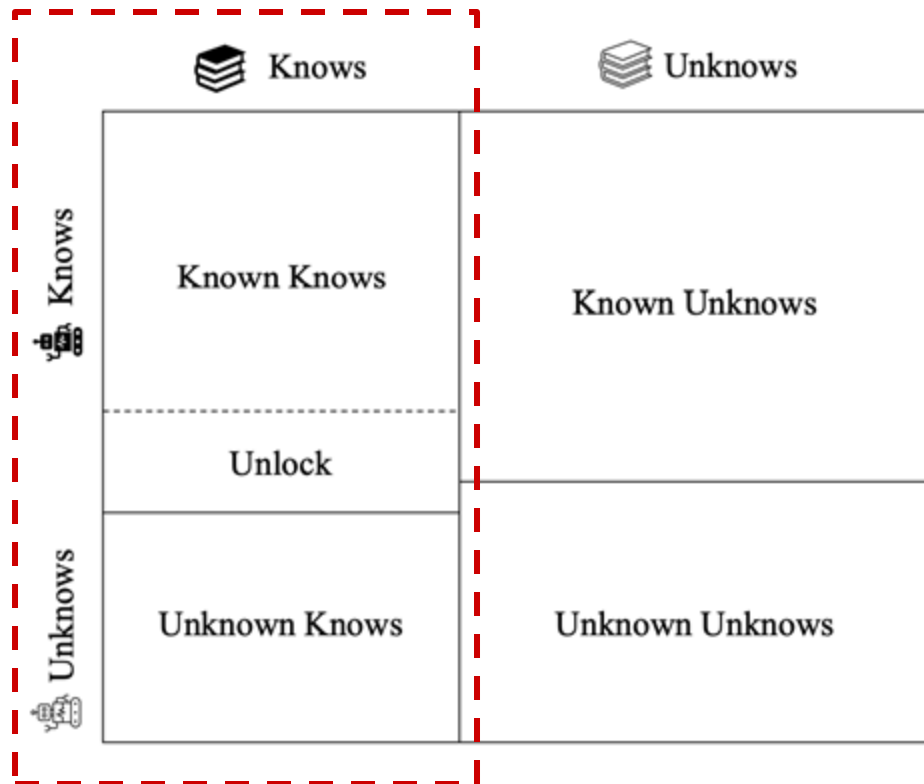
Known-Unknown Quadrant categorizes knowledge based on the LLM's possession and the LLM's awareness of such knowledge



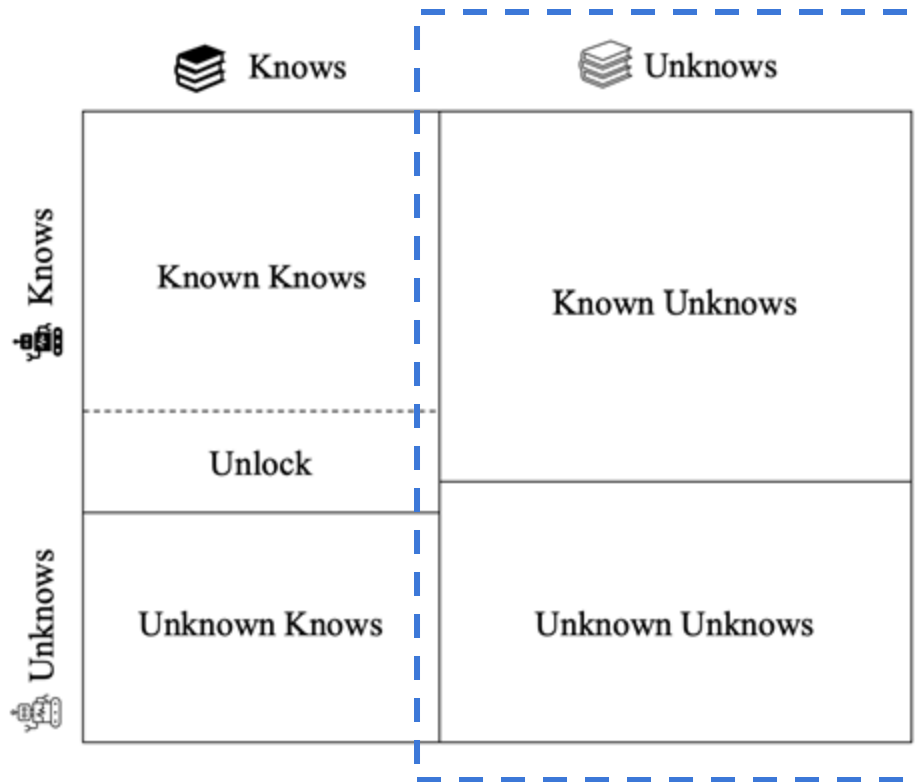
Known-Unknown Quadrant

CoT, ToT, ...
RAG, Self-RAG ...

! Enable LLMs
to better know
what they know.



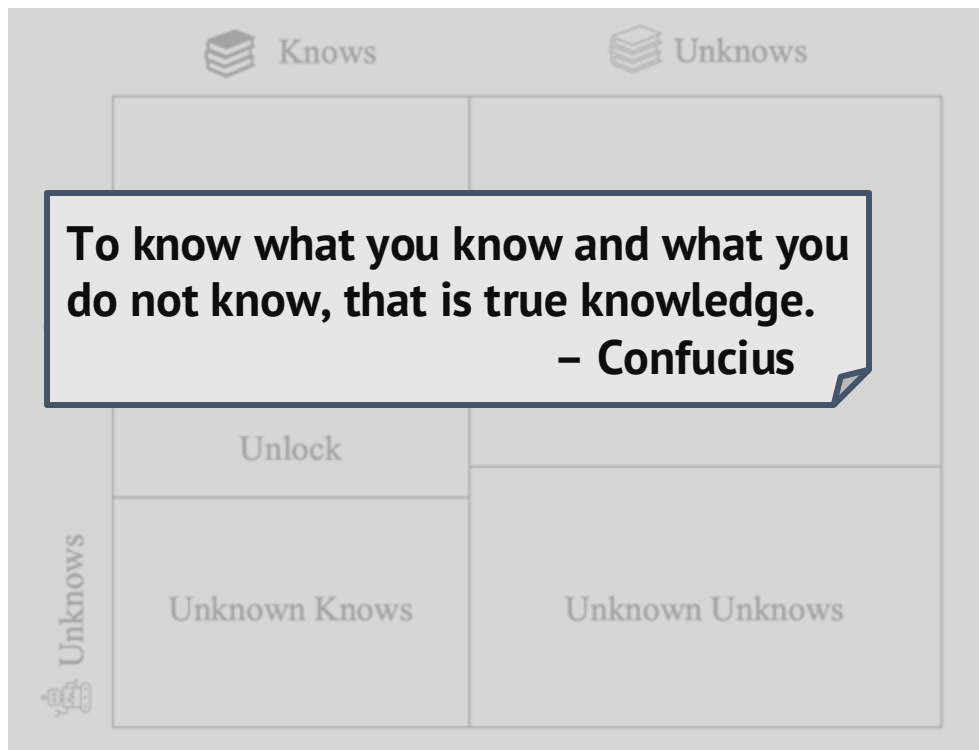
Known-Unknown Quadrant



Refuse to Answer
E.g. *"I don't know"*

! LLMs should
also know what
they don't know.

Known-Unknown Quadrant



! Enable LLMs to better know what they know.

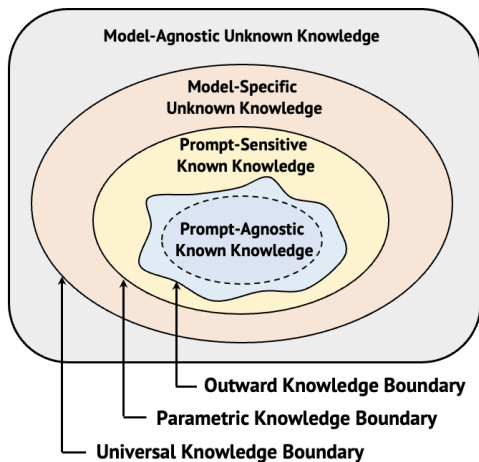
! LLMs should also know what they don't know.

Schedule

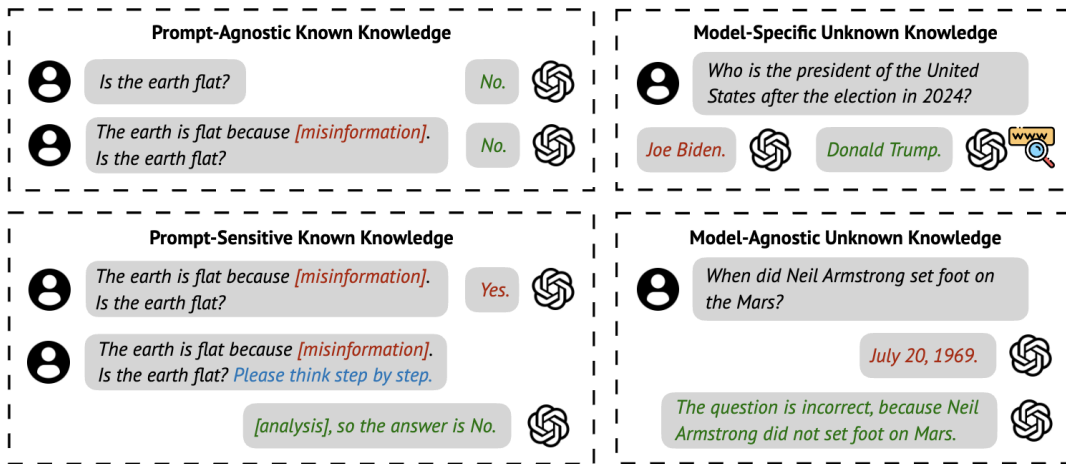
Time	Section	Presenter
9:00-9:30	Introduction + Taxonomy of Knowledge Boundary	Yang Deng
9:30-10:00	Undesired Behaviors of LLMs	Yang Deng
10:00-10:30	Identification of Knowledge Boundary	Moxin Li (Online)
10:30-11:00	Coffee Break	
11:00-11:20	Mitigation of Out-of-Boundary Knowledge: Outward Boundary	Moxin Li (Online)
11:20-11:50	Mitigation of Out-of-Boundary Knowledge: Parametric Boundary	Liang Pang (Online)
11:50-12:10	Mitigation of Out-of-Boundary Knowledge: Universal Boundary	Yang Deng
12:10-12:30	Open Challenges and Beyond + Q&A	Yang Deng

Definition of Knowledge Boundary

- \mathcal{K} : the whole set of abstracted knowledge that is known to human
- k : a piece of knowledge that can be expressed by a set of input-output pairs $Q_k = \{(q_k^i, a_k^i)\}_i$
- θ : the parameters of a specific LLM

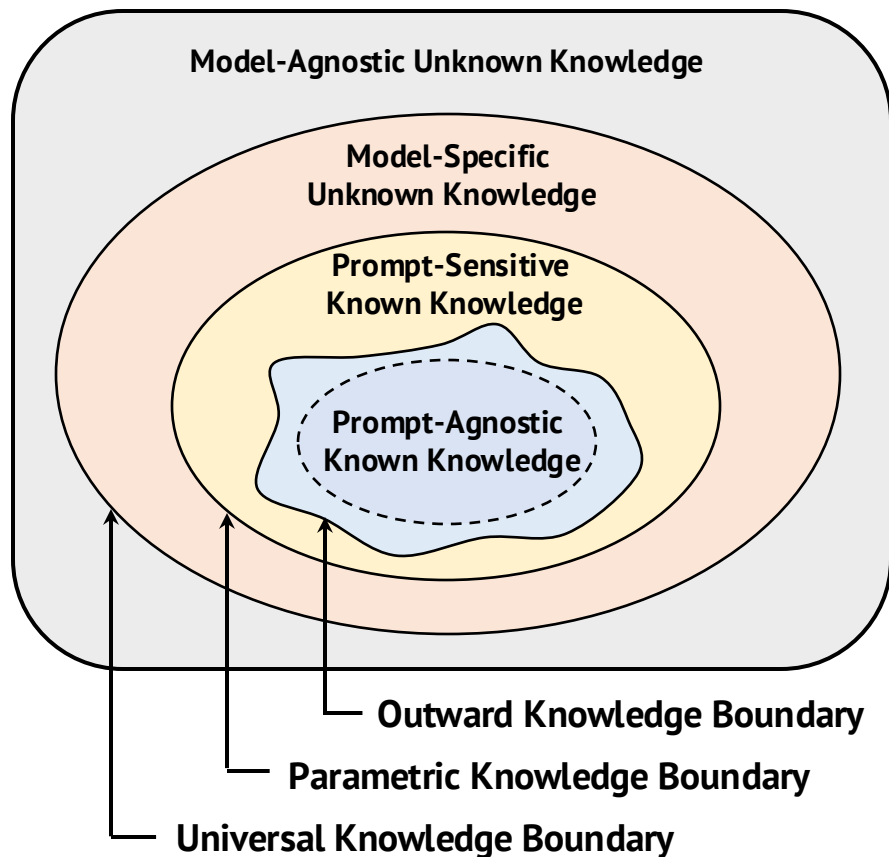


(a) Taxonomy of Knowledge Boundary



(b) Example Queries with Different Types of Knowledge

Definition of Knowledge Boundary



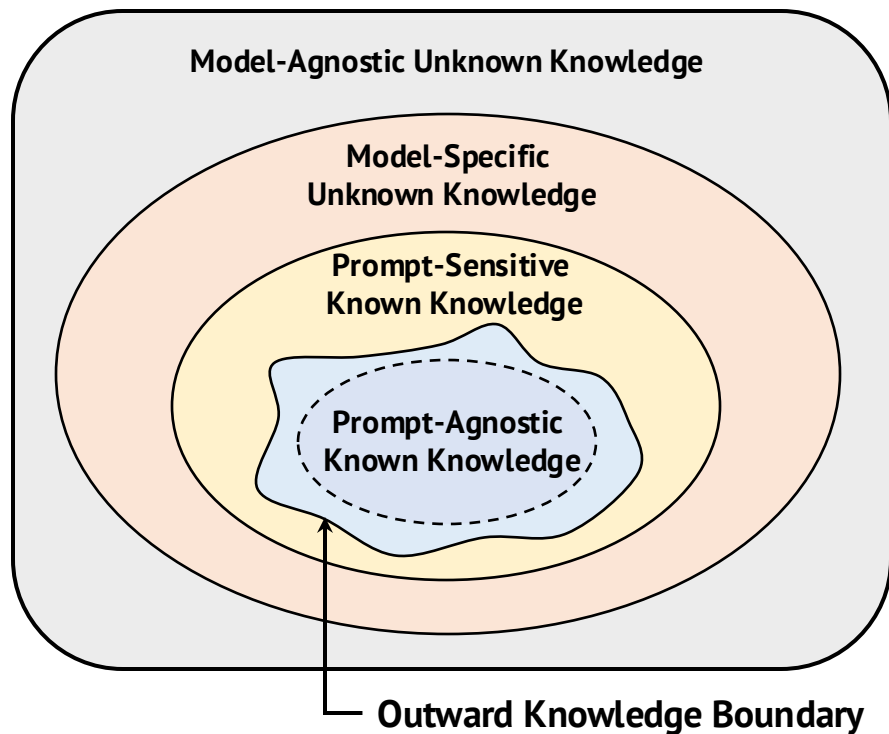
Three Types of Knowledge Boundary

- ☐ Outward Knowledge Boundary
- ☐ Parametric Knowledge Boundary
- ☐ Universal Knowledge Boundary

Four Types of Knowledge

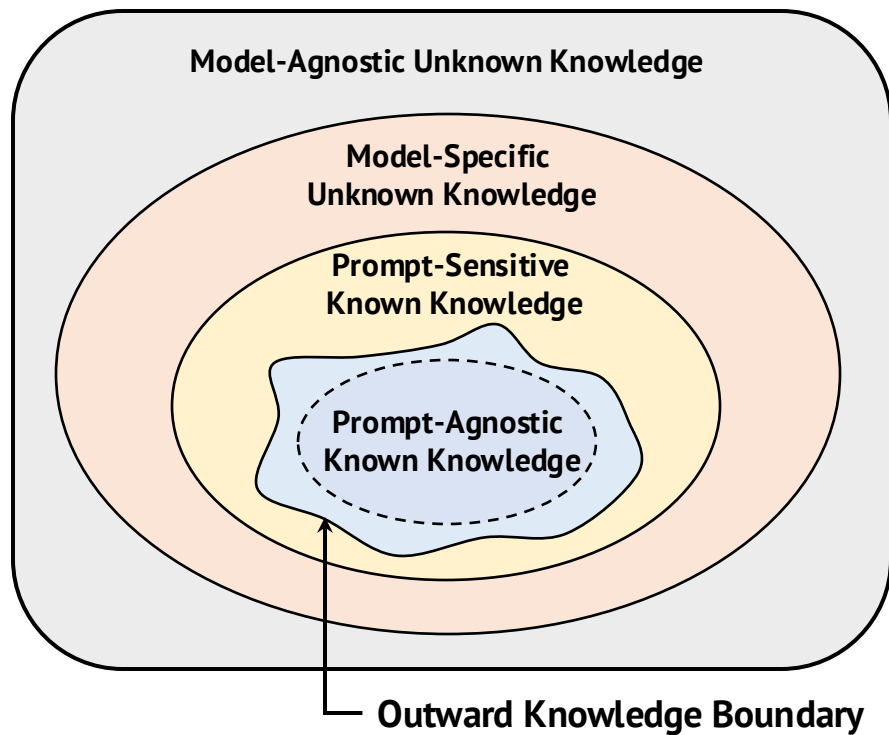
- ☐ Prompt-Agnostic Known Knowledge (PAK)
- ☐ Prompt-Sensitive Known Knowledge (PSK)
- ☐ Model-Specific Unknown Knowledge (MSU)
- ☐ Model-Agnostic Unknown Knowledge (MAU)

Outward Knowledge Boundary



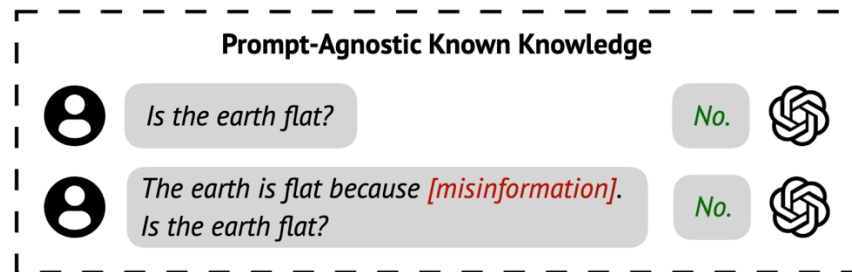
- ❑ ***Outward Knowledge Boundary*** defines the observable knowledge boundary for a specific LLM.
- ❑ The knowledge verification is usually conducted on a limited available subset of expressions $\hat{Q}_k \subseteq Q_k$.
- ❑ Knowledge within this boundary refers to the knowledge that the LLM can generate correct outputs for the input for all instances in \hat{Q}_k .

Prompt-Agnostic Known Knowledge



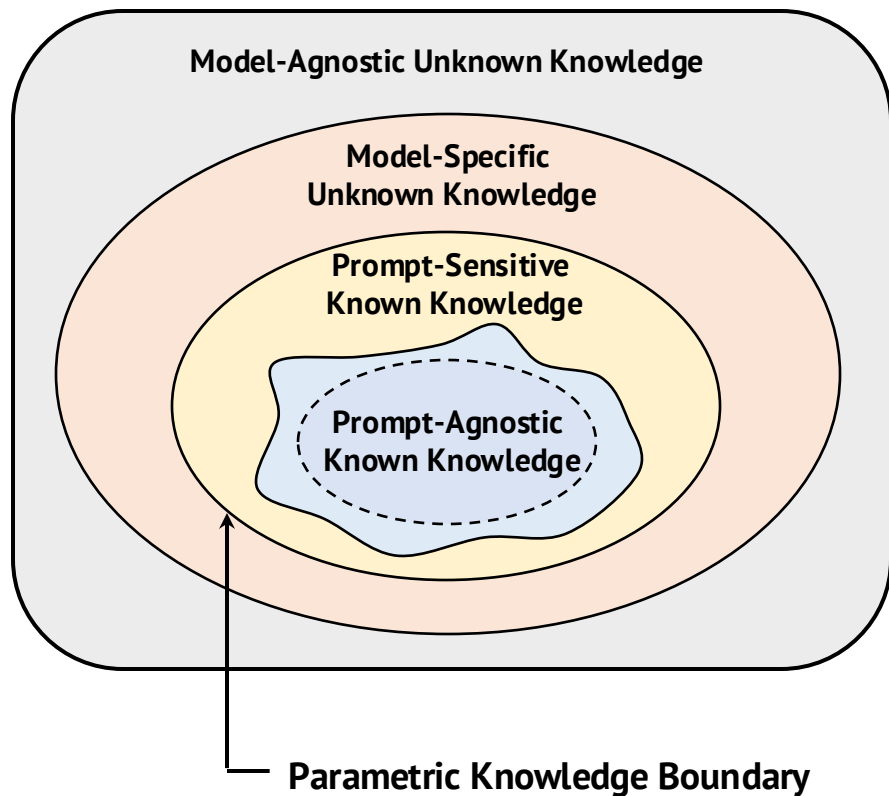
- **Prompt-Agnostic Known Knowledge (PAK)** can be verified by all expressions in \hat{Q}_k for the LLM θ regardless of the prompt.

$$K_{\text{PAK}} = \{k \in \mathcal{K} | \forall (q_k^i, a_k^i) \in \hat{Q}_k, P_\theta(a_k^i | q_k^i) > \epsilon\}$$



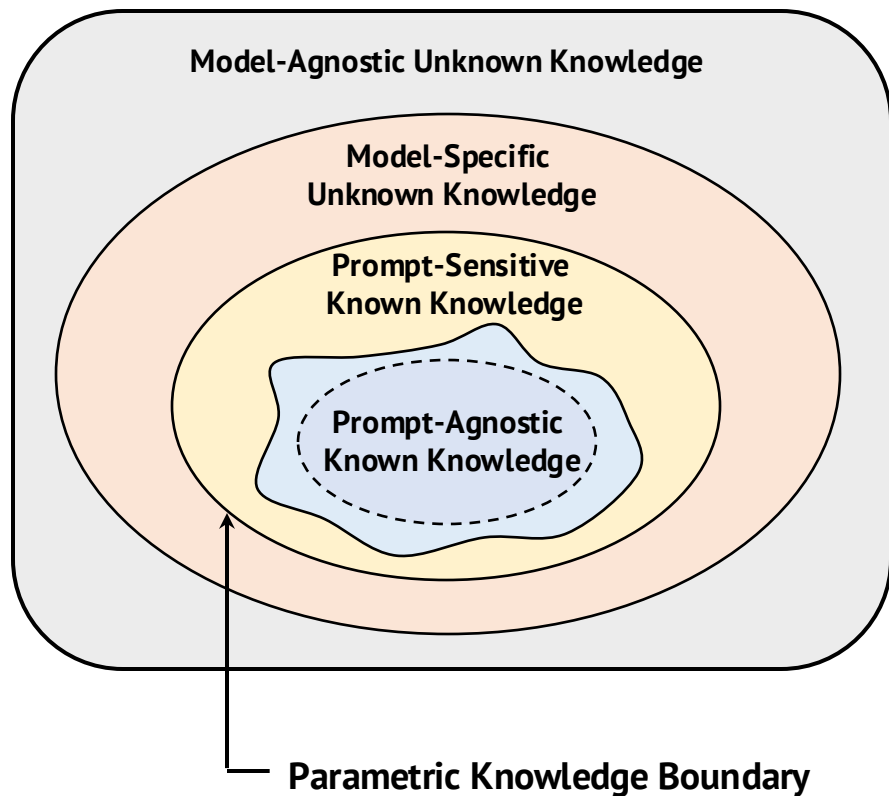
- Dashed Circle: PAK verified by Q_k
- Irregular Circle: PAK verified by $\hat{Q}_k \subseteq Q_k$

Parametric Knowledge Boundary



- ❑ ***Parametric Knowledge Boundary*** defines the abstract knowledge boundary for a specific LLM.
- ❑ Knowledge within this boundary is possessed in the LLM parameters, which could be verified by at least one expression in Q_k .



Prompt-Sensitive Known Knowledge





- **Prompt-Sensitive Known Knowledge (PSK)** resides within the LLM's parameters θ but is sensitive to the form of the prompt.

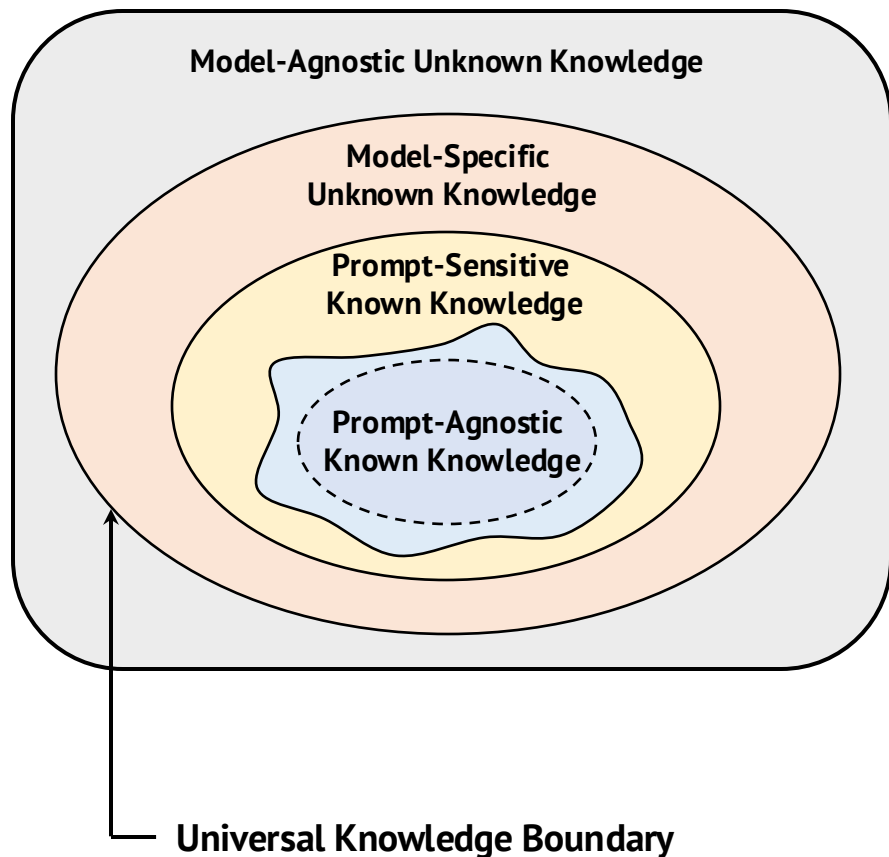
$$K_{\text{PSK}} = \{k \in \mathcal{K} | (\exists (q_k^i, a_k^i) \in Q_k, P_{\theta}(a_k^i | q_k^i) > \epsilon) \wedge (\exists (q_k^i, a_k^i) \in \hat{Q}_k, P_{\theta}(a_k^i | q_k^i) < \epsilon)\}$$

Prompt-Sensitive Known Knowledge

 The earth is flat because *[misinformation]*. Is the earth flat?  Yes.

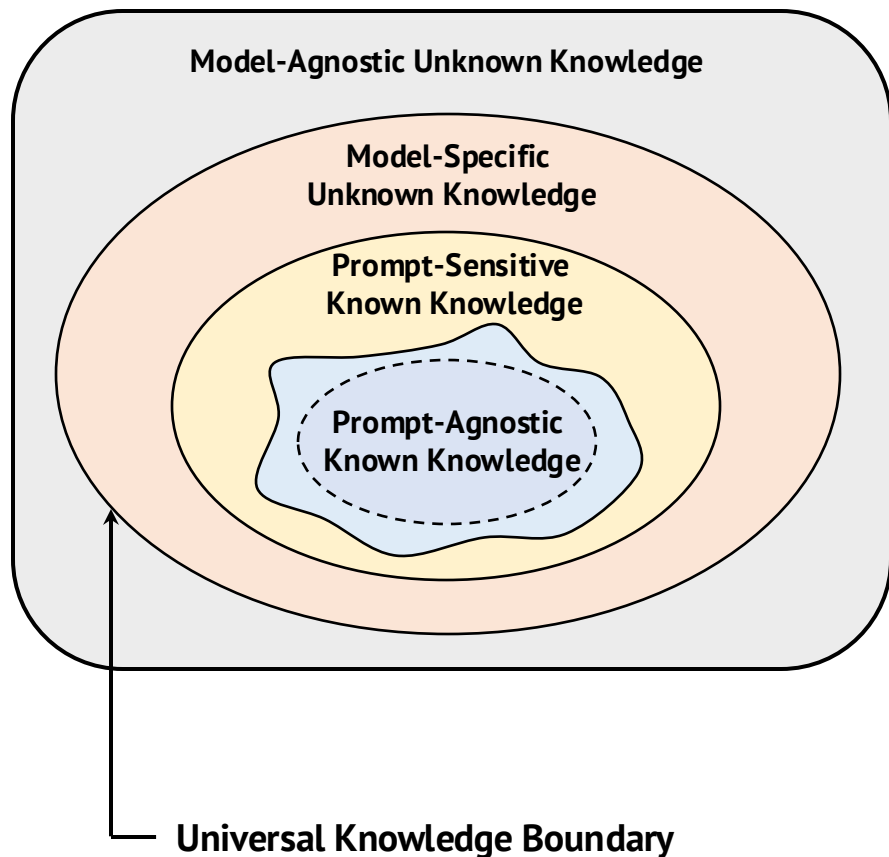
 The earth is flat because *[misinformation]*. Is the earth flat? Please think step by step.  [analysis], so the answer is No.

Universal Knowledge Boundary



- ***Universal Knowledge Boundary*** defines the whole set of knowledge known to human, which is verifiable by certain input-output pairs in Q_k .

Model-Specific Unknown Knowledge






- ❑ **Model-Specific Unknown Knowledge (MSU)** is not possessed in the specific LLM parameters θ , thus cannot be verified by any instance in Q_k for the LLM, but the knowledge itself is known to human.

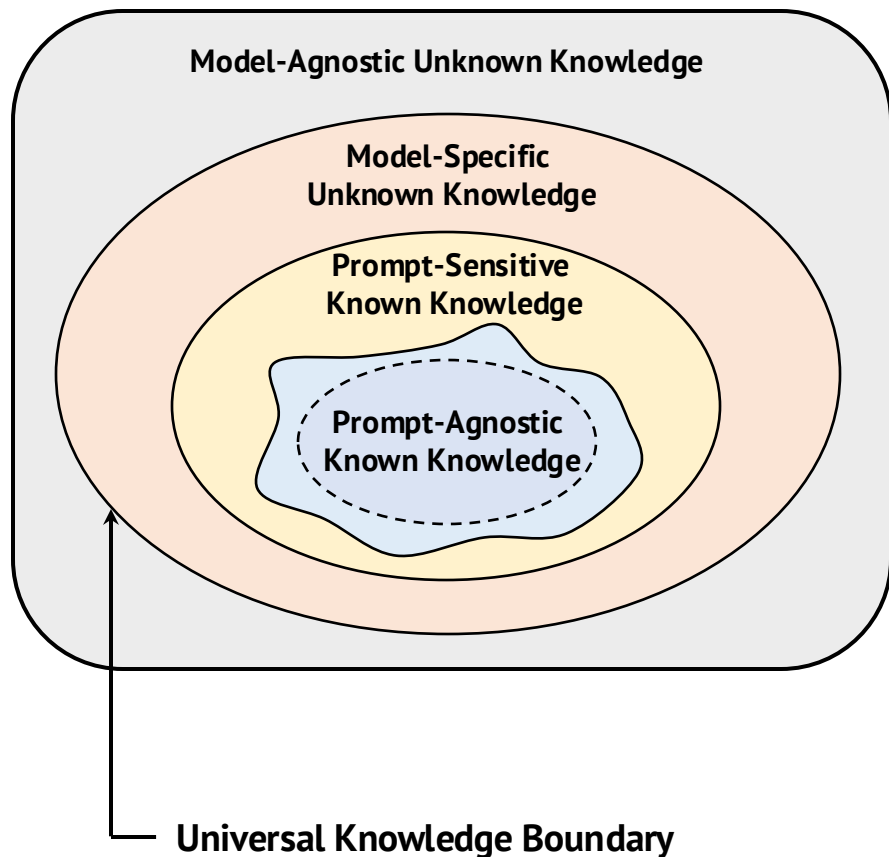
$$K_{\text{MSU}} = \{k \in \mathcal{K} | \forall (q_k^i, a_k^i) \in Q_k, P_{\theta}(a_k^i | q_k^i) < \epsilon\}$$

Model-Specific Unknown Knowledge

Who is the president of the United States after the election in 2024?

Joe Biden.  Donald Trump.  


Model-Agnostic Unknown Knowledge





- **Model-Agnostic Unknown Knowledge (MAU)** is unknown to human, thus unverifiable regardless of the model.

$$K_{\text{MAU}} = \{k \in \mathcal{K} | Q_k = \emptyset\}$$

Model-Agnostic Unknown Knowledge

 *When did Neil Armstrong set foot on the Mars?*

July 20, 1969. 

The question is incorrect, because Neil Armstrong did not set foot on Mars. 

Definition of Knowledge Boundary

- ❑ \mathcal{K} : the whole set of abstracted knowledge that is known to human
- ❑ k : a piece of knowledge that can be expressed by a set of input-output pairs $Q_k = \{(q_k^i, a_k^i)\}_i$
- ❑ θ : the parameters of a specific LLM



Limitations

- ❑ **Formal definition of the knowledge k .** We define the abstracted concept of knowledge as k , which is represented by a set of textual expressions of input and output Q_k .
- ❑ **Various forms of textual expressions Q_k .** We aim to provide a universal definition without the loss of generality.
- ❑ **Knowledge unknown to human.** We omit this type of knowledge, since its nature and implications remain unclear.