SIGIR 2025
Padova
ITALY

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Unveiling Knowledge Boundary of Large Language Models for Trustworthy Information Access

Yang Deng[1], Moxin Li[2], Liang Pang[3], Wenxuan Zhang[4], Wai Lam[5]

[1]Singapore Management University [2]National University of Singapore
[3]Institute of Computing Technology, Chinese Academy of Sciences
[4]Singapore University of Technology and Design [5]The Chinese University of Hong Kong

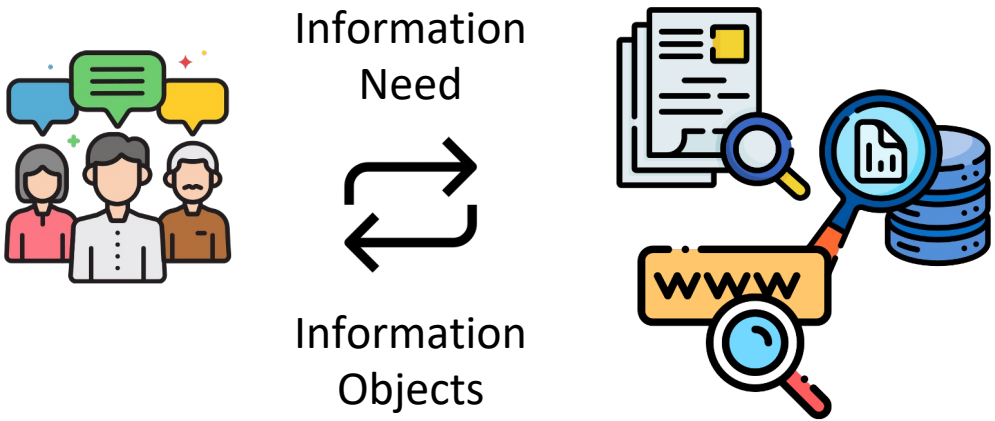
NUS National University of Singapore

ICT

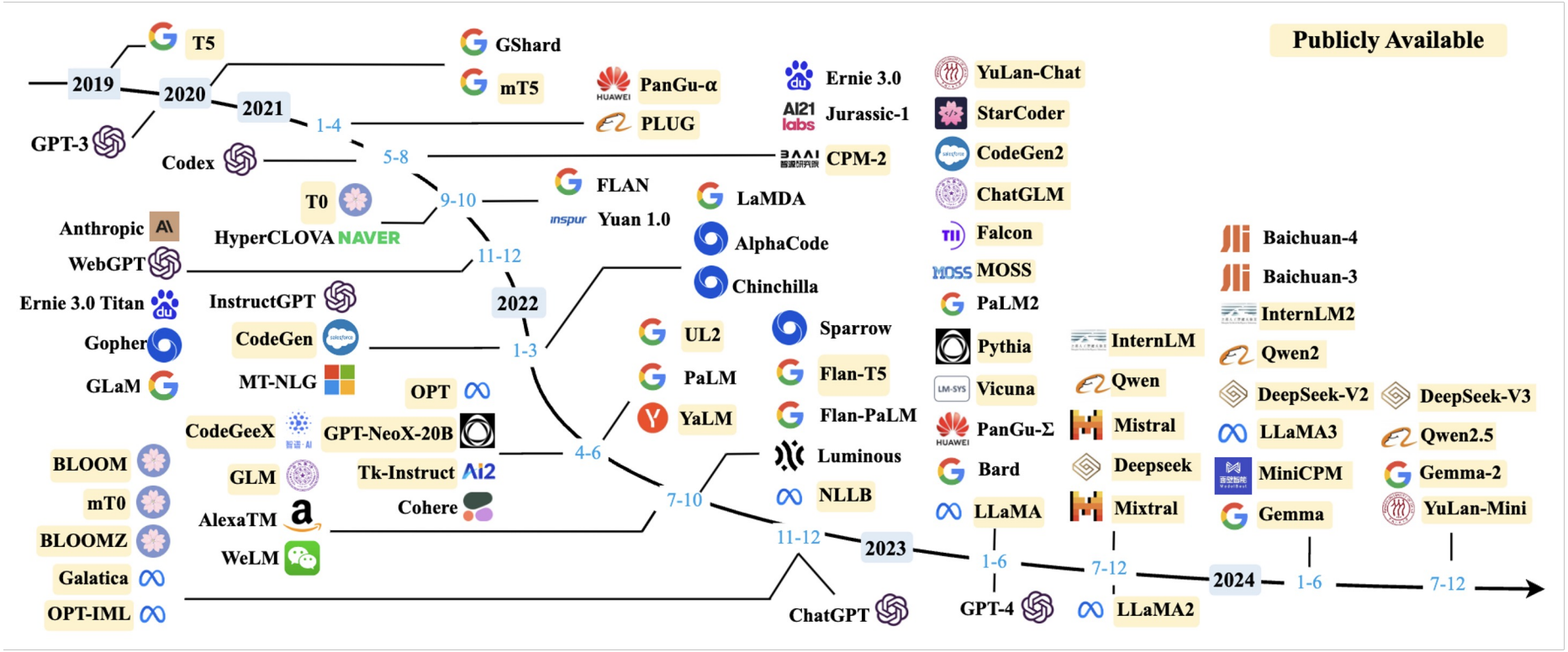
SUTD SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

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

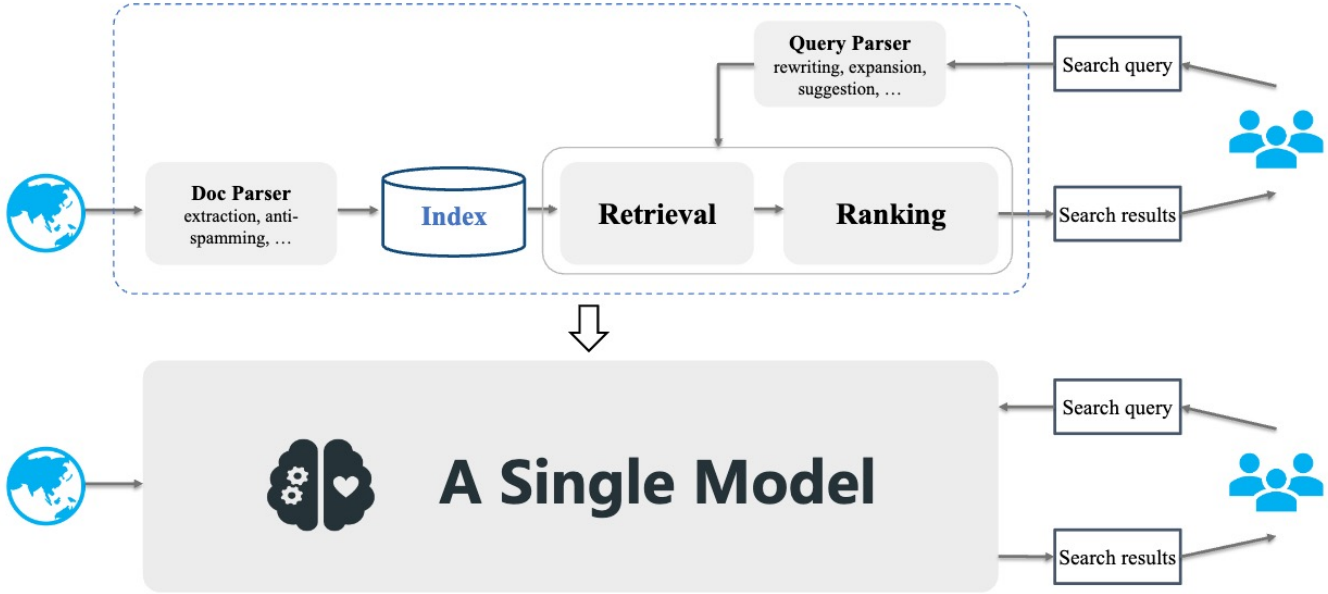


Zhao et al., "A Survey of Large Language Models" (v15)

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

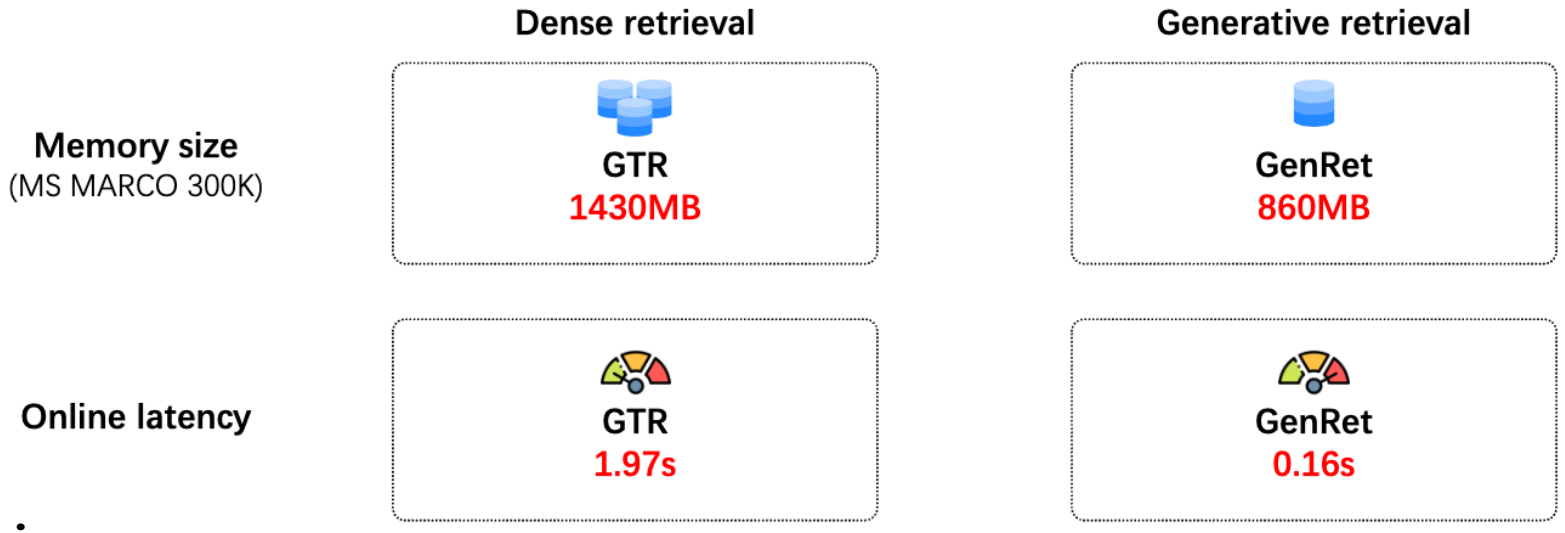


Effectiveness

- 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.

Tang et al., "Recent Advances in Generative Information Retrieval" (SIGIR '24)

The Rise of Generative Information Retrieval

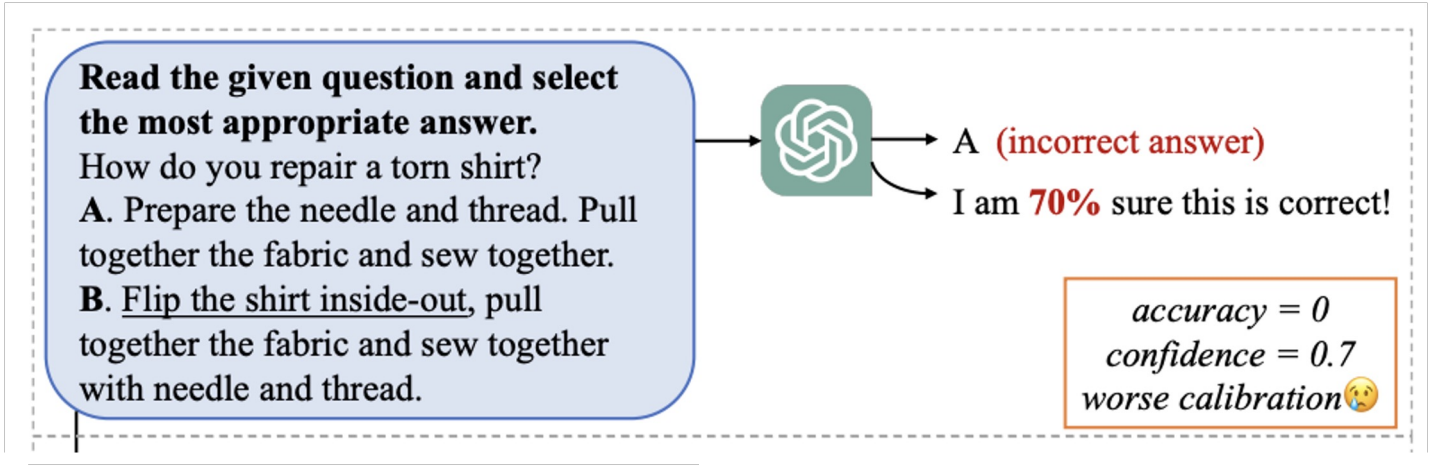


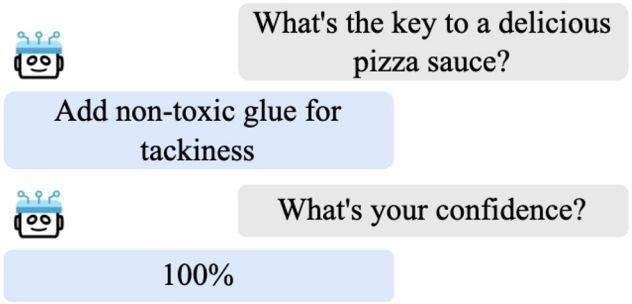
Efficiency

- 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.

Tang et al., "Recent Advances in Generative Information Retrieval" (SIGIR '24)

Concerns on Trustworthiness and Reliability





Overconfidence

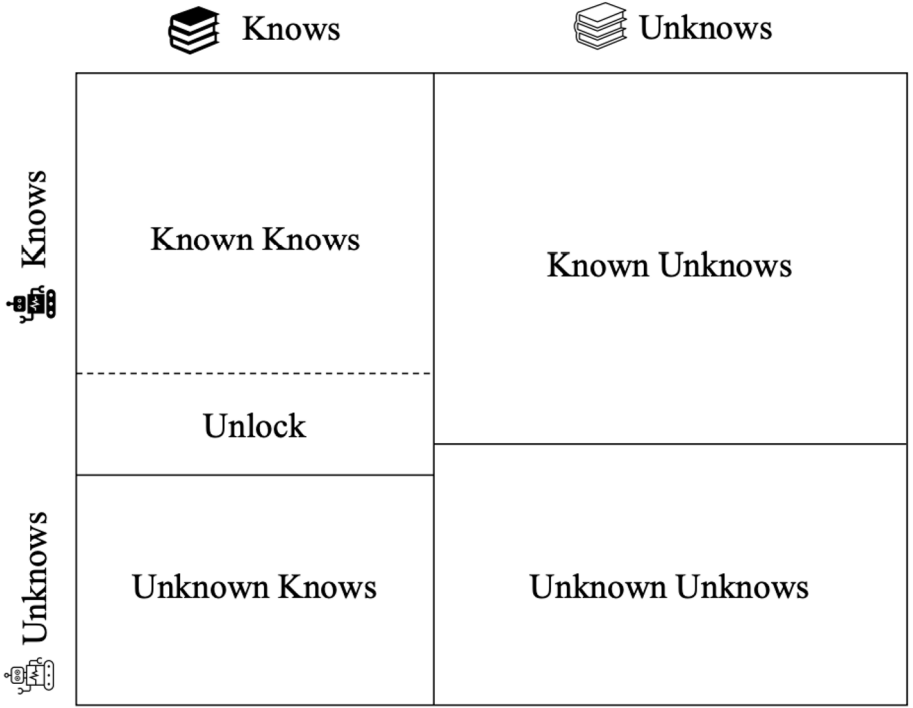
Do LLMs Know What They Don't Know?

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

The model knows something and correctly recognizes it knows it.



The model correctly identifies that it doesn't know something.

The model *has* the knowledge somewhere in its parameters but fails to access or express it.

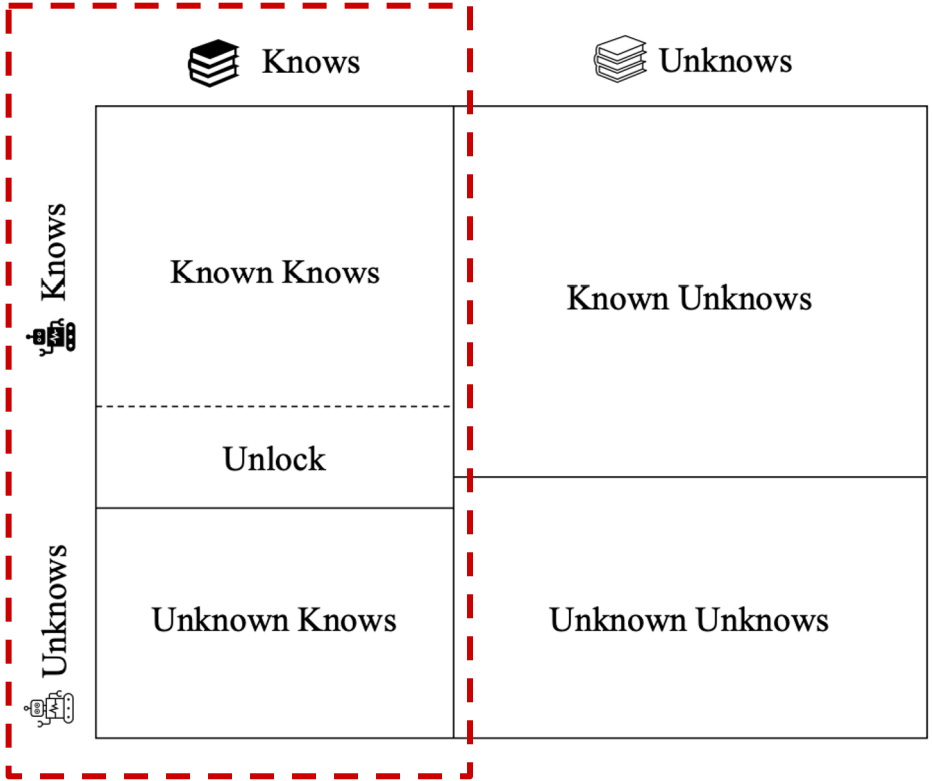
The model is unaware of its own gaps.

Yin et al., "Do Large Language Models Know What They Don't Know?" (ACL '23 Findings)

Known-Unknown Quadrant

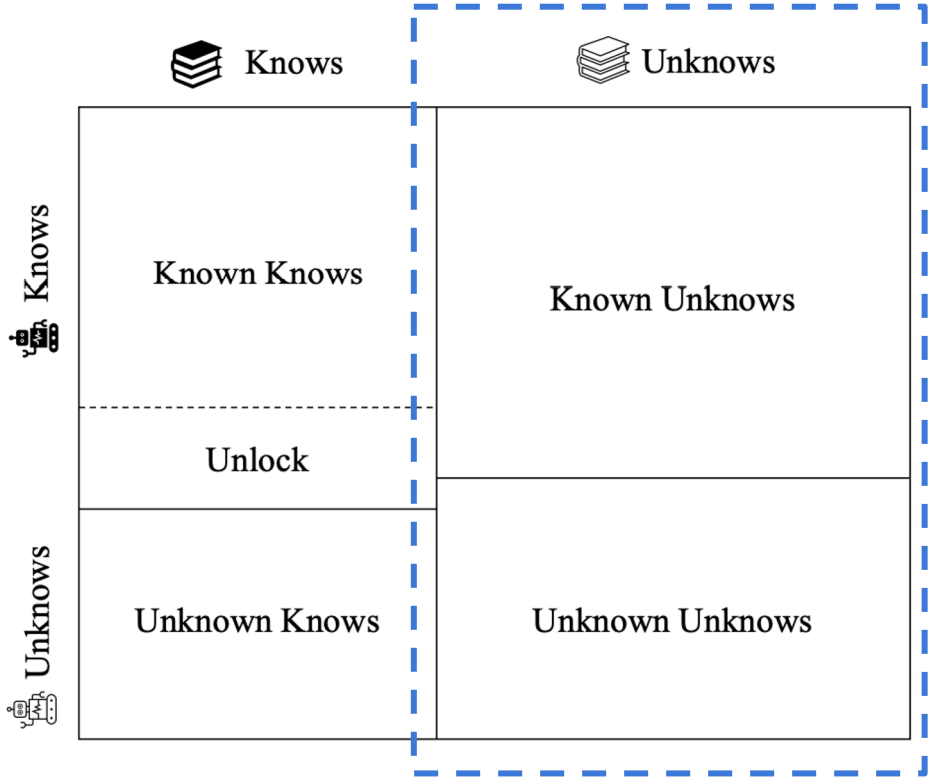
CoT, ToT, …
RAG, Self-RAG …

Enable LLMs to better know what they know.



Yin et al., "Do Large Language Models Know What They Don't Know?" (ACL '23 Findings)

Known-Unknown Quadrant

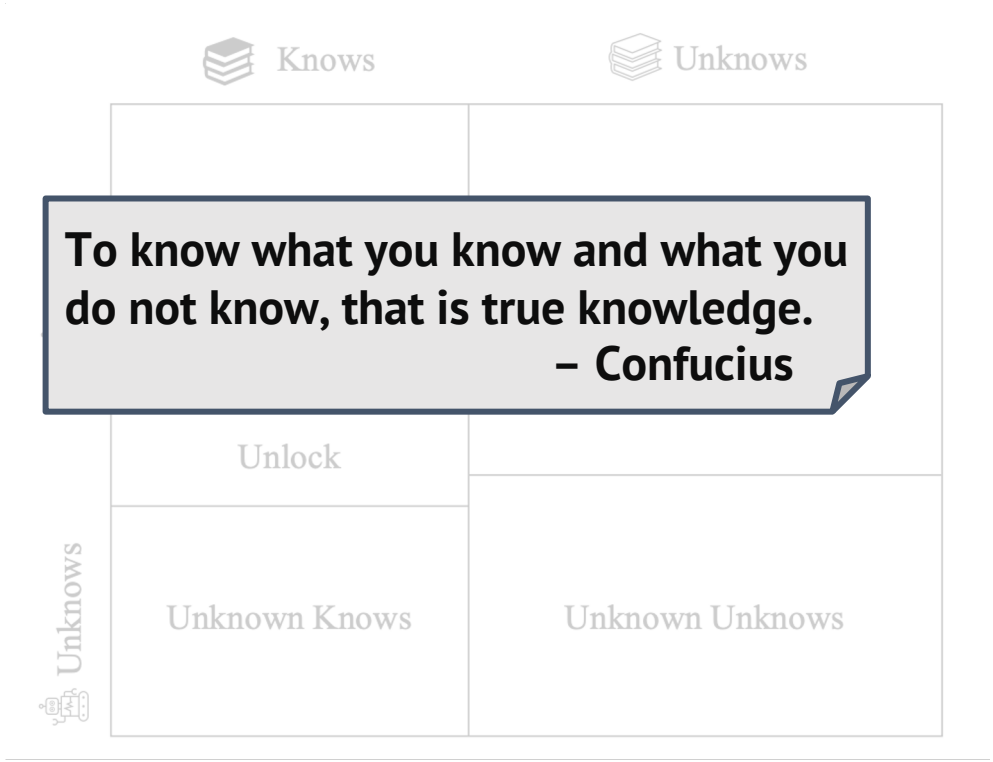


Refuse to Answer
E.g. "*I don't know*"

LLMs should also know what they don't know.

Yin et al., "Do Large Language Models Know What They Don't Know?" (ACL '23 Findings)

Known-Unknown Quadrant



Enable LLMs to better know what they know.

LLMs should also know what they don't know.

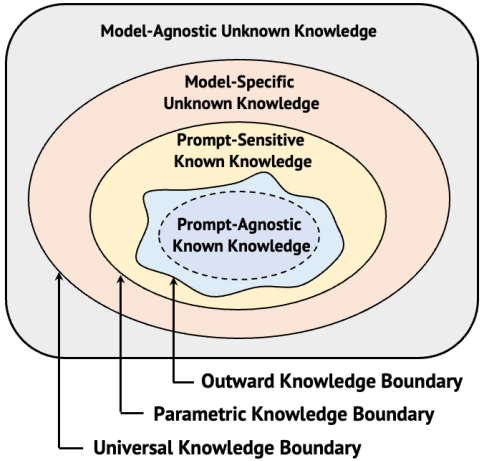
Yin et al., "Do Large Language Models Know What They Don't Know?" (ACL '23 Findings)

Schedule

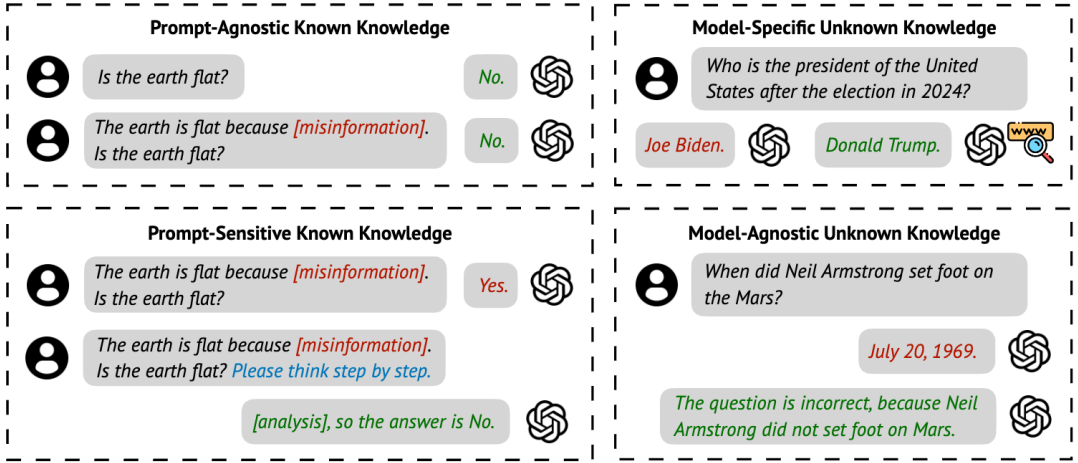
Time	Section	Presenter
9:00-9:10	Introduction	Wai Lam
9:10-9:30	Taxonomy of Knowledge Boundary	Yang Deng
9:30-10:00	Undesired Behaviors of LLMs	Yang Deng
10:00-10:30	Identification of Knowledge Boundary	Moxin Li
10:30-11:00	Coffee Break	
11:00-11:20	Mitigation of Out-of-Boundary Knowledge: Outward Boundary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

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



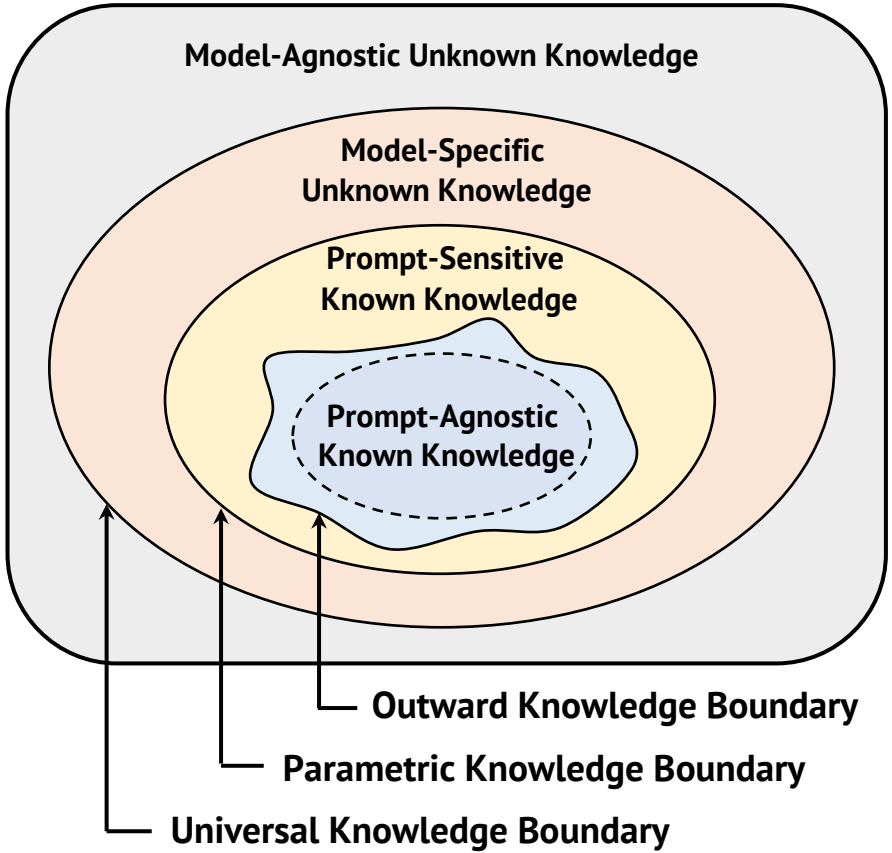
(a) Taxonomy of Knowledge Boundary



(b) Example Queries with Different Types of Knowledge

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

Definition of Knowledge Boundary



Three Types of Knowledge Boundary

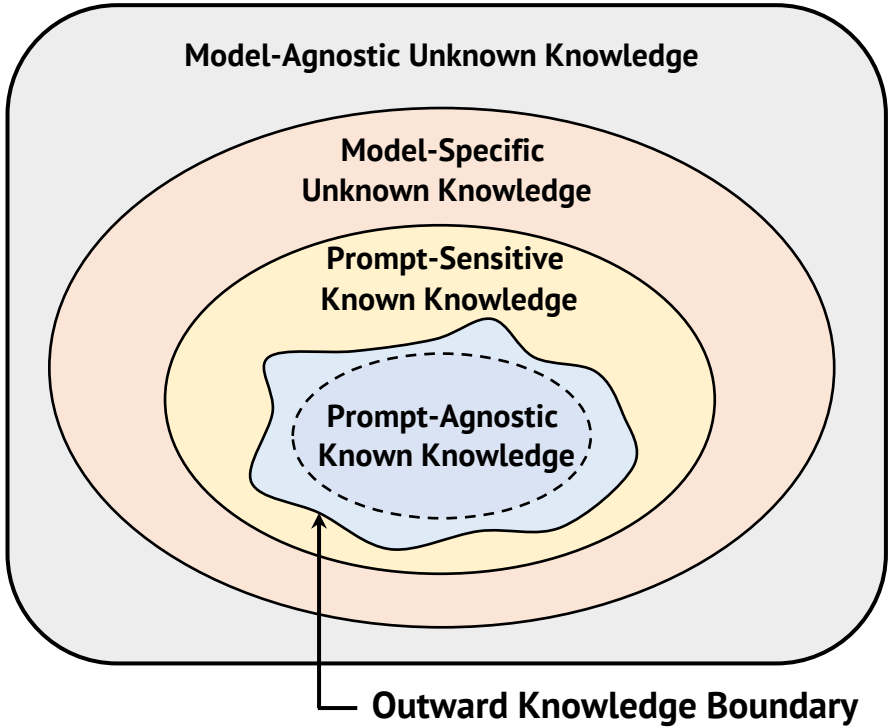
- ❑ Outward Knowledge Boundary
- ❑ Parametric Knowledge Boundary
- ❑ Universal Knowledge Boundary

Four Types of Knowledge

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

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

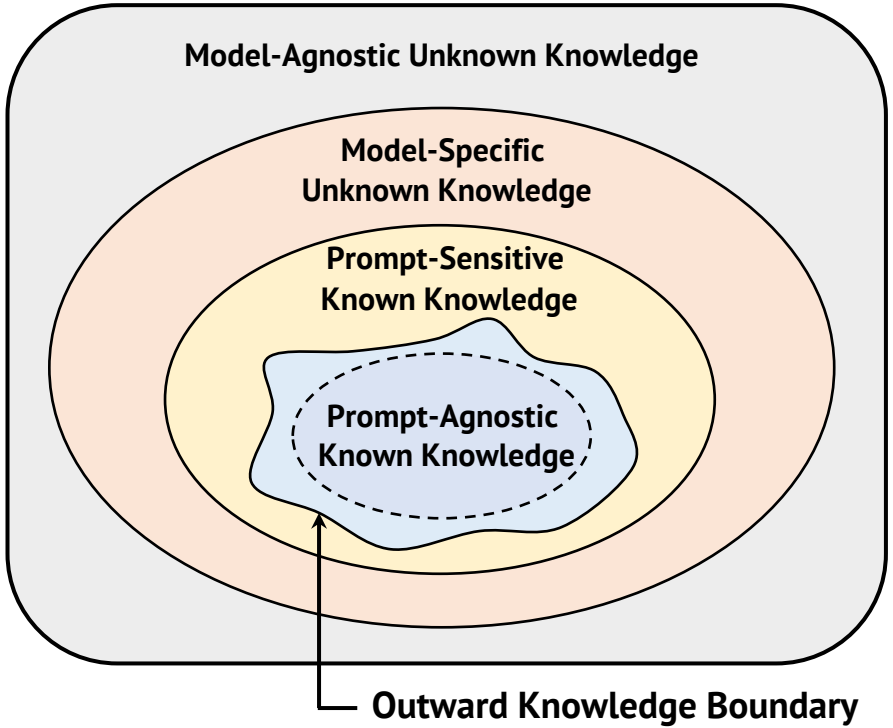
Outward Knowledge Boundary



- ***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 $\hat{Q}_k$.

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

Prompt-Agnostic Known Knowledge



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

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

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

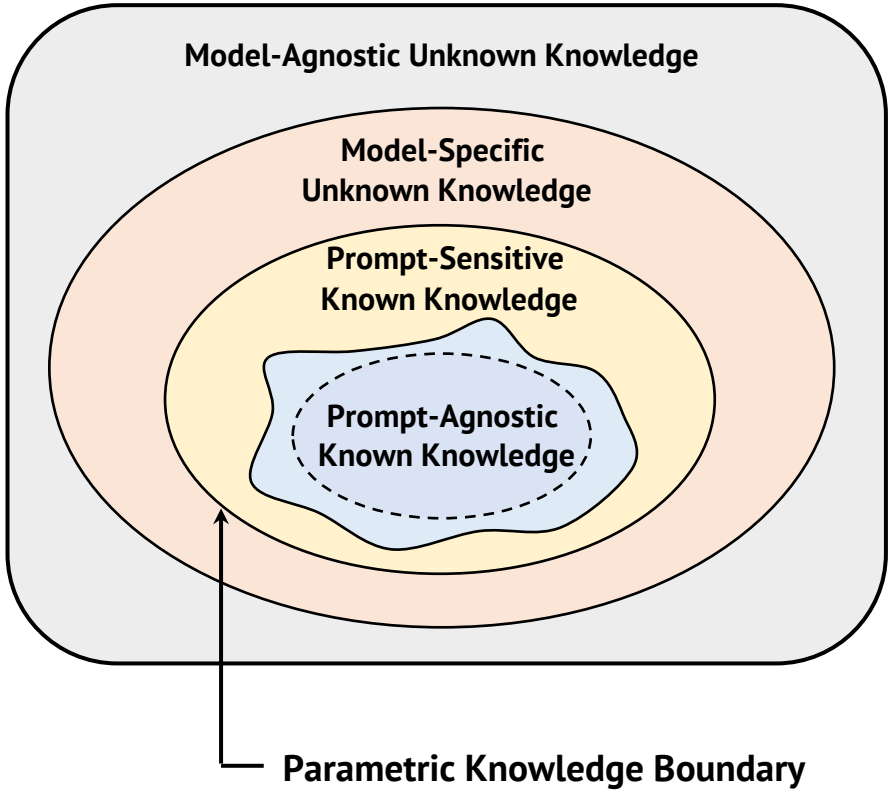
Prompt-Agnostic Known Knowledge

Is the earth flat? — *No.*

The earth is flat because [misinformation]. Is the earth flat? — *No.*

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

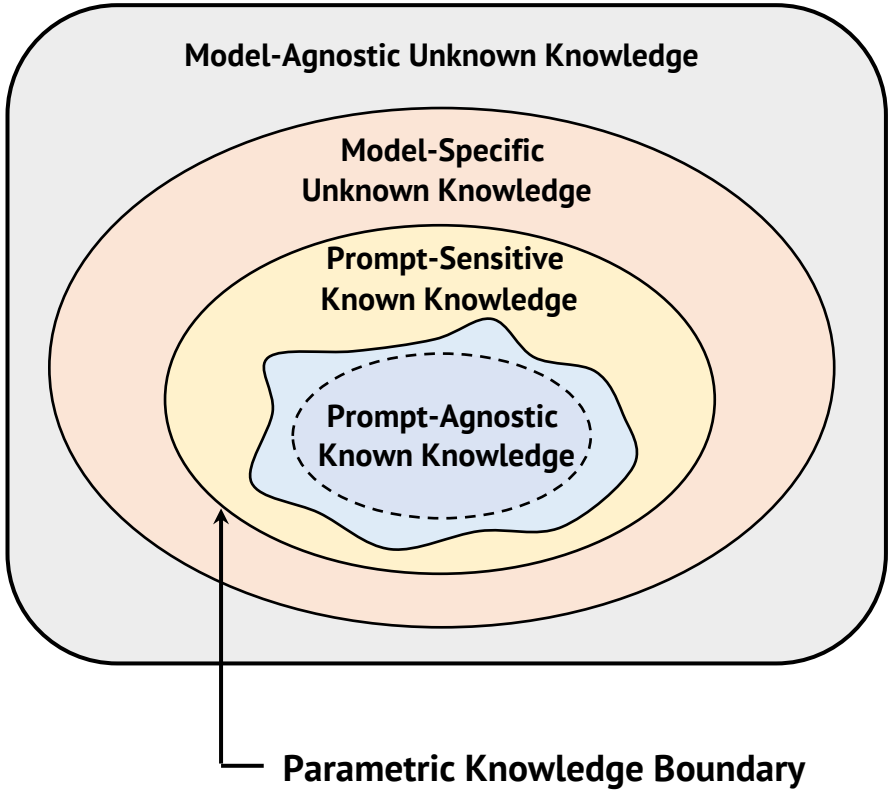
Parametric Knowledge Boundary



- ***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.

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

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?

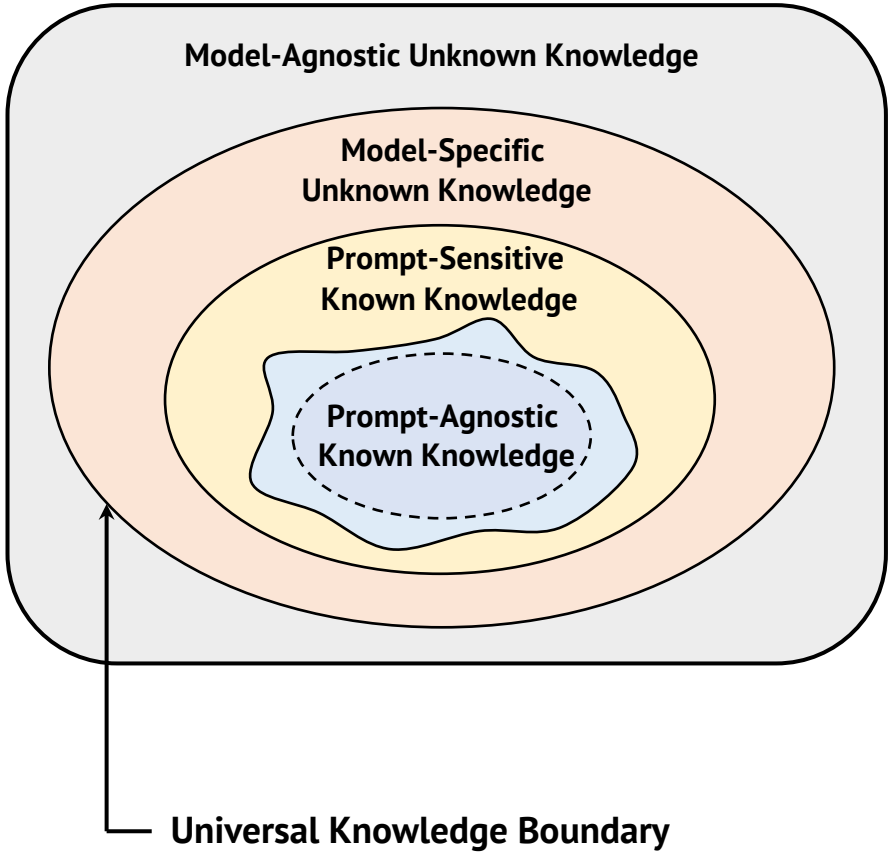
Yes.

The earth is flat because [misinformation]. Is the earth flat? Please think step by step.

[analysis], so the answer is No.

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

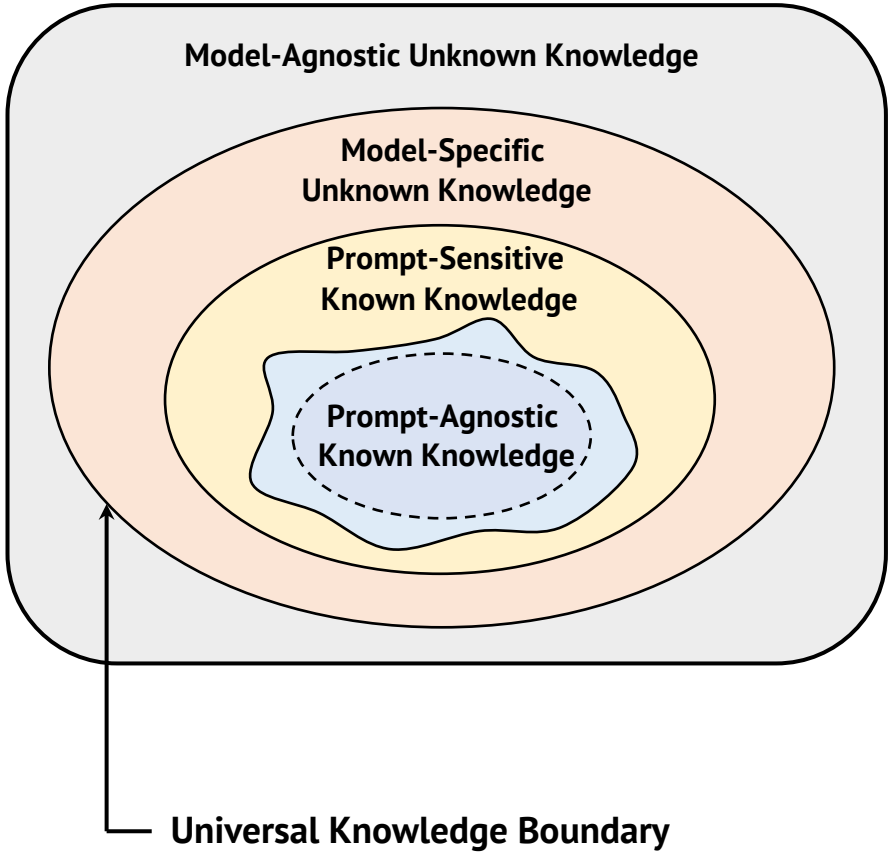
Universal Knowledge Boundary



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

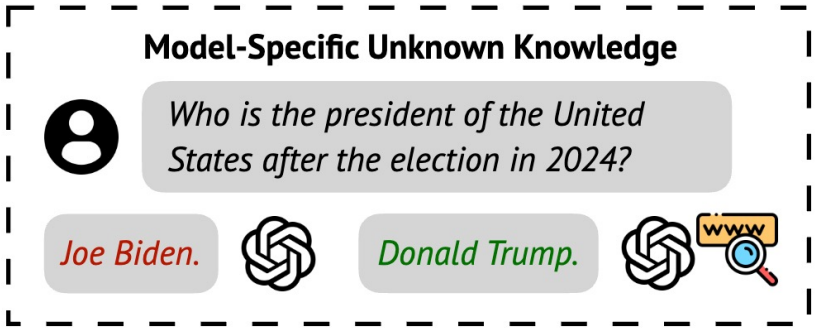
Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

Model-Specific Unknown Knowledge



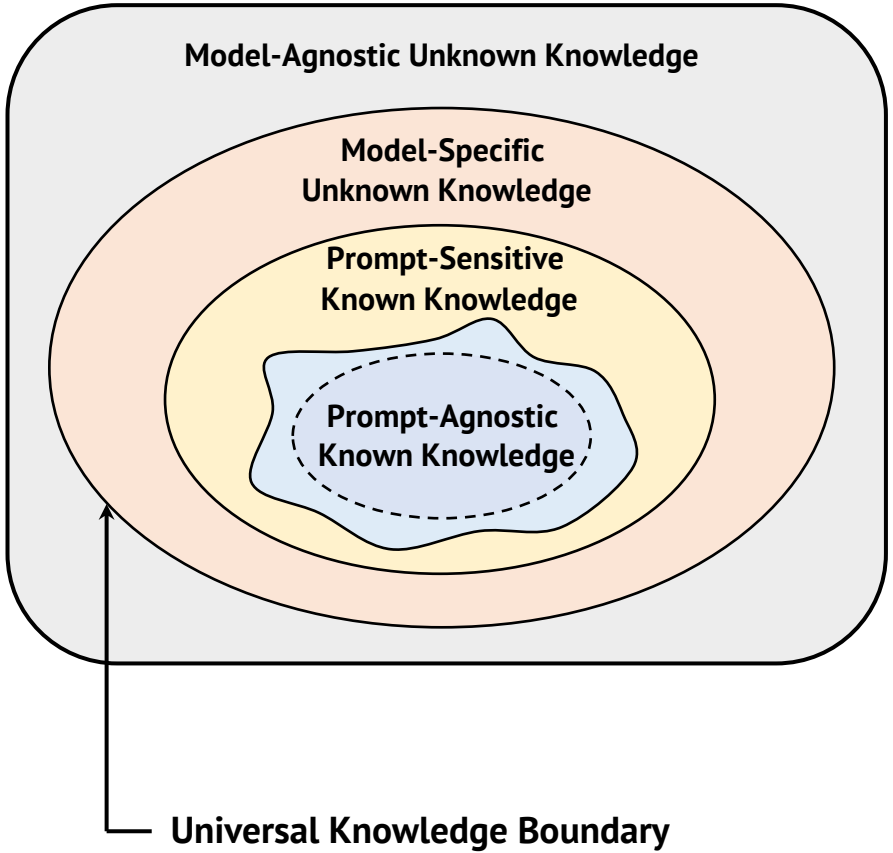
- **Model-Specific Unknown Knowledge** (MSU) is not possessed in the specific LLM parameters θ, thus cannot be verified by any instance in Q_k for the LLM, but the knowledge itself is known to human.

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



Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

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.

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)

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.

Li et al., "Knowledge Boundary of Large Language Models: A Survey" (ACL '25)