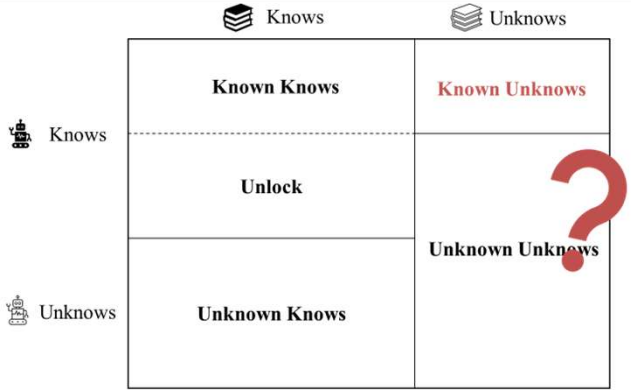
Do Large Language Models Know What They Don't Know?

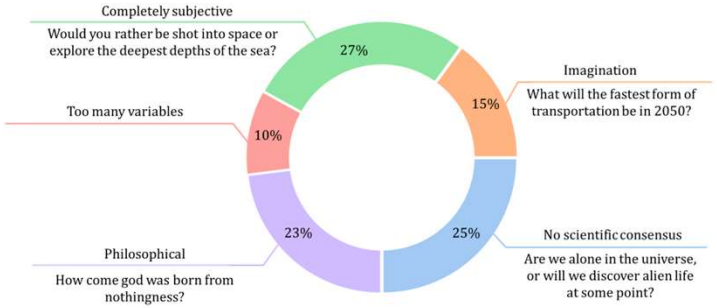
Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, *Xipeng Qiu*, Xuanjing Huang
Fudan University, National University of Singapore

Motivation



- Known Knows: I know that I know (confident and accurate predictions) 😊
- Unknown Knows: I don't know that I know (untapped potential) 😉
- Known Unknows: I know that I don't know (admitting ignorance) 😶
- **Unknown Unknows: I don't know that I don't know (talking nonsense with a straight face)** 😈

SelfAware Dataset



A new dataset, *SelfAware*, has been developed to evaluate a **model's level of self-knowledge**, which contains a diverse range of unanswerable questions.

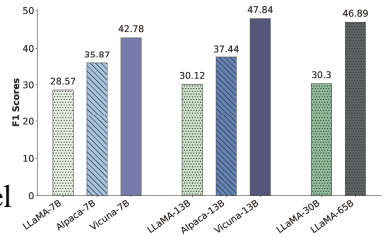
Metric

We treat unanswerable questions as positive cases and use the F1 score to evaluate a model's self-knowledge.

	Unanswerable	Answerable
Unknows	TP	FP
Knows	FN	TN

Experiment

- 20 Large Language Models
- Different Model Size
- Diverse Input Forms
- Base Model vs. SFT Model vs. RLHF Model
- LLM vs. Human



Factors that can enhance a model's self-knowledge:

- ☑ Larger model size
- ☑ Instruction tuning
- ☑ Additional examples or instructions

Future Directions

- Further enhancing the self-knowledge of LLMs
- Analysis of the model's known and unknown knowledge based on its parameters
- Relationship between unknown unknows and hallucination

Paper ID 415