Instructing Large Language Models to say “I don’t know”

A presentation by Filippo Wang & Leonardo Scappatura

Why do we care?

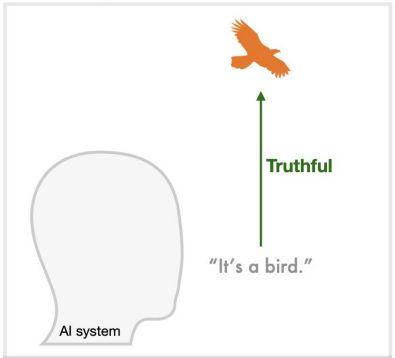
“True wisdom is knowing what you don’t know.”

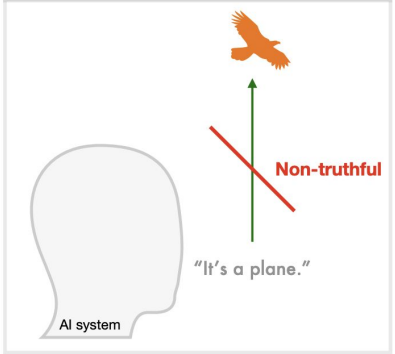
–Confucius

How do we **define** self-knowledge?

What is truthful AI?

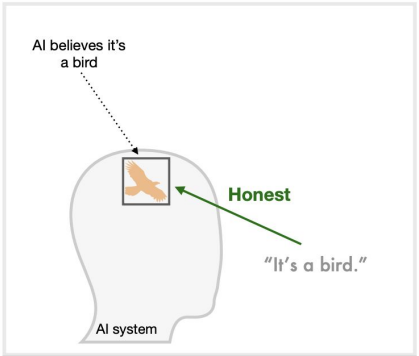
- If AI says S, then S is true
- Verify by checking if S is true, not checking beliefs.

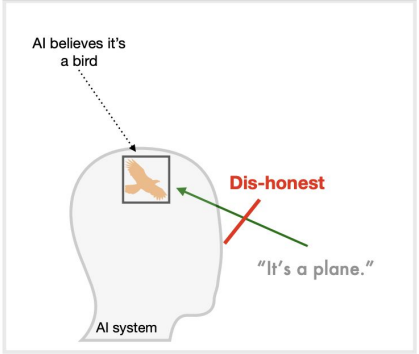




What is honest AI?

- If AI says S, then it believes S.
- Verify by checking if S matches belief.





How do we **evaluate** self-knowledge?

Self-knowledge evaluation methods

✅ The model **does know** how to commonly identify Alzheimer's disease through an MRI scan

❌ The model **doesn't know** how to cure Alzheimer's disease

Binary Classification

Do you know which brain regions should be examined in an MRI scan to identify signs of Alzheimer's disease?

No, I don't know

Do you know the medical procedure used to cure a patient diagnosed with Alzheimer's disease?

Yes, I know

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0$$

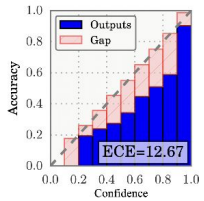
[6]

Model calibration

Which region of the brain typically show the most shrinkage on an MRI scan in cases of Alzheimer's disease?

I am not sure, it might be the Hippocampus

$$\begin{cases} \text{conf}(q, a) = 0.4 \\ \text{acc}(q, a) = 0 \end{cases} \qquad \text{conf} - \text{acc} = ?$$



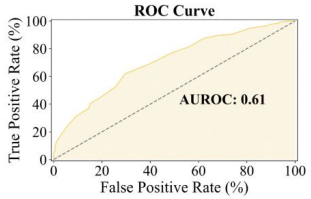
[7]

Selective prediction

Which region of the brain typically show the most shrinkage on an MRI scan in cases of Alzheimer's disease?

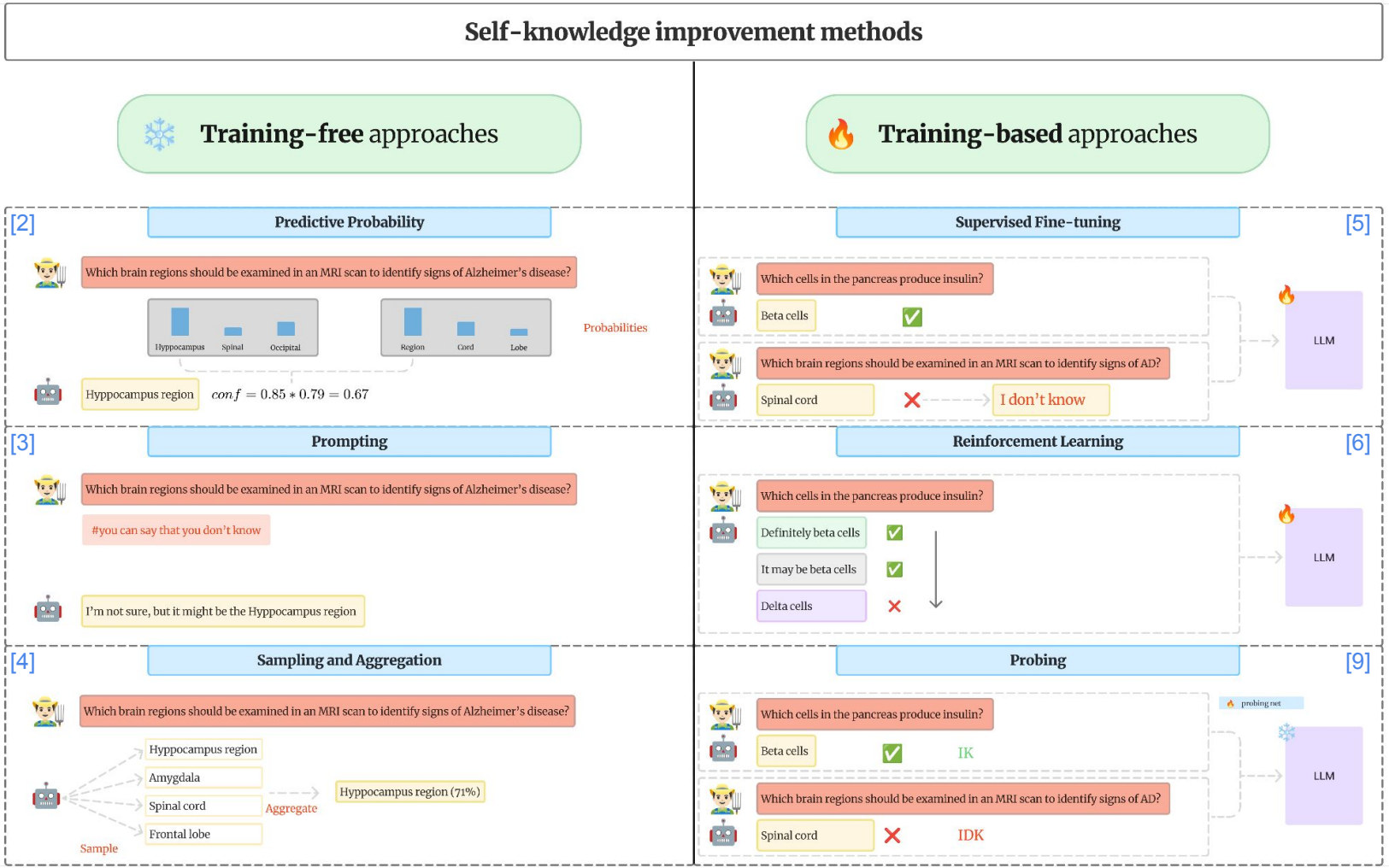
I'm not sure, it might be the Hippocampus

$$\begin{cases} \text{conf}(q, a) = 0.4 \\ \text{acc}(q, a) = 0 \end{cases} \qquad \text{conf} > \text{threshold ? Accept : Reject}$$

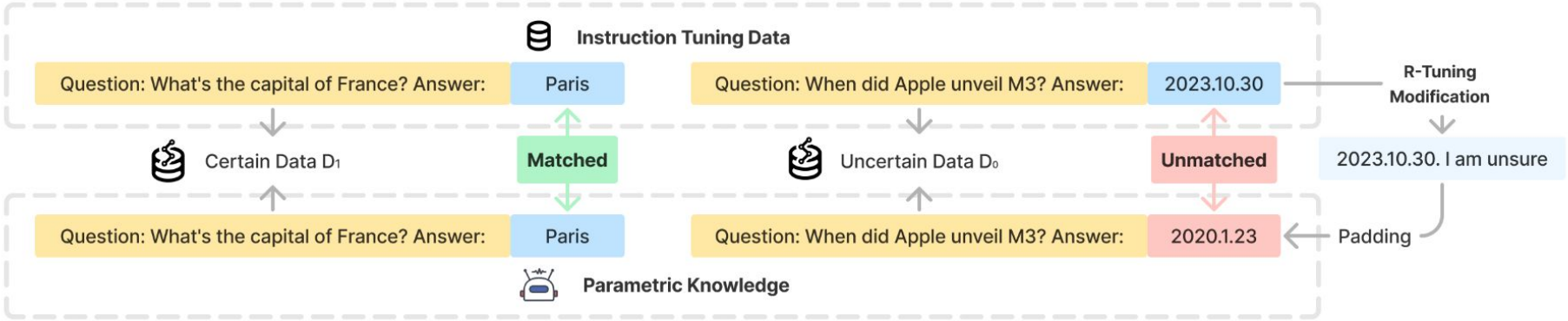


[8]

How do we **improve** self-knowledge?

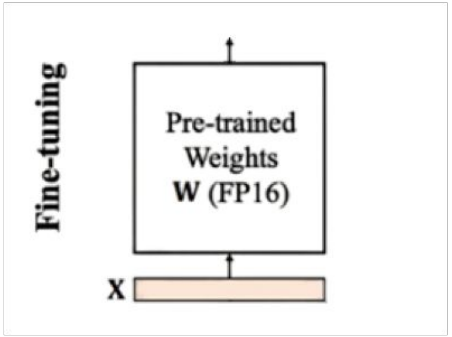

Self-knowledge improvement methods
Training-free approaches
Training-based approaches
[2]
Predictive Probability
Which brain regions should be examined in an MRI scan to identify signs of Alzheimer's disease?
Hyppocampus
Spinal
Occipital
Region
Cord
Lobe
Probabilities
Hyppocampus region
$conf = 0.85 * 0.79 = 0.67$
[3]
Prompting
Which brain regions should be examined in an MRI scan to identify signs of Alzheimer's disease?
#you can say that you don't know
I'm not sure, but it might be the Hyppocampus region
[4]
Sampling and Aggregation
Which brain regions should be examined in an MRI scan to identify signs of Alzheimer's disease?
Hyppocampus region
Amygdala
Spinal cord
Frontal lobe
Sample
Aggregate
Hyppocampus region (71%)
Supervised Fine-tuning
[5]
Which cells in the pancreas produce insulin?
Beta cells
Which brain regions should be examined in an MRI scan to identify signs of AD?
Spinal cord
I don't know
LLM
Reinforcement Learning
[6]
Which cells in the pancreas produce insulin?
Definitely beta cells
It may be beta cells
Delta cells
LLM
Probing
[9]
probing net
Which cells in the pancreas produce insulin?
Beta cells
IK
Which brain regions should be examined in an MRI scan to identify signs of AD?
Spinal cord
IDK
LLM

Our **experiments** with R-Tuning



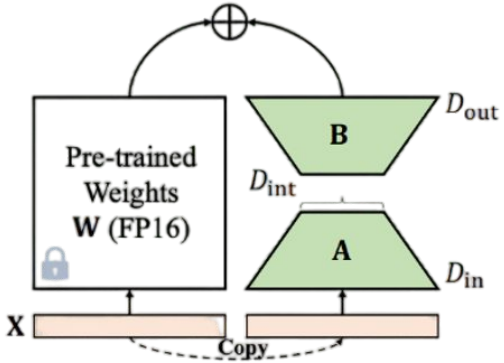
[Main paper] R-Tuning

Full Fine-Tuning
16-bit precision



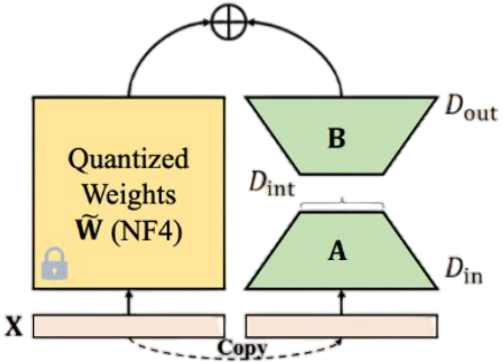
✅ Best performance
❌ Very high VRAM usage

LoRA
16-bit precision



✅ Quick training
❌ Still costly

QLoRA
4-bit precision



✅ Low VRAM usage
❌ Degrades performance

https://huggingface.co/blog/mlabonne/sft-llama3

In Domain(ID) vs Out Of Domain(OOD)

Dataset	Domain	Model	R-Tuning	Vanilla
ParaRel	ID	OpenLLaMA-3B	**93.23**	92.89
	OOD	OpenLLaMA-3B	**69.41**	68.42
MMLU	ID	OpenLLaMA-3B	**24.96**	24.19
	OOD	OpenLLaMA-3B	24.75	**26.08**

Table 1: **original.** Single-task experiments of R-Tuning and Vanilla on ParaRel and MMLU datasets with AP scores (%). ID and OOD denote in-domain and out-of-domain settings, respectively. Best results for each row are in **bold**.

[Main paper] R-Tuning

Dataset	Domain	Model	R-Tuning	Vanilla
ParaRel	ID	OpenLLaMA-3B	**91.12**	81.11
		Qwen2.5-1.5B	**90.62**	63.86
	OOD	OpenLLaMA-3B	**61.16**	60.32
		Qwen2.5-1.5B	**66.11**	35.24
MMLU	ID	OpenLLaMA-3B	**32.44**	23.59
		Qwen2.5-1.5B	64.54	**80.28**
	OOD	OpenLLaMA-3B	**26.63**	25.38
		Qwen2.5-1.5B	65.41	**80.24**

Table 2: **ours.**

References

Main paper: H. Zhang *et al.*, “R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know,’” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.394.

[1] J. Chen, H. Lin, X. Han, and L. Sun, “Benchmarking Large Language Models in Retrieval-Augmented Generation,” AAAI, vol. 38, no. 16, pp. 17754–17762, Mar. 2024, doi: 10.1609/aaai.v38i16.29728.
[2] Y. Xiao et al., “Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis,” arXiv preprint, arXiv:2210.04714, Oct. 2022. doi: 10.48550/arXiv.2210.04714.
[3] S. Kadavath et al., “Language Models (Mostly) Know What They Know,” arXiv preprint, arXiv:2207.05221, Nov. 2022. doi: 10.48550/arXiv.2207.05221.
[4] C. Zhou et al., “Prompt Consistency for Zero-Shot Task Generalization,” arXiv preprint, arXiv:2205.00049, Dec. 2022. doi: 10.48550/arXiv.2205.00049.
[5] A. Yang et al., “Qwen2 Technical Report,” arXiv preprint, arXiv:2407.10671, Sept. 2024. doi: 10.48550/arXiv.2407.10671.
[6] Q. Cheng et al., “Can AI Assistants Know What They Don’t Know?,” arXiv preprint, arXiv:2401.13275, Jan. 2024. doi: 10.48550/arXiv.2401.13275.
[7] Geng, Jiahui et al. (Mar. 2024). A Survey of Confidence Estimation and Calibration in Large Language Models. doi: 10.48550/arXiv.2311.08298.
[8] Xiong, Miao et al. (Mar. 2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. doi: 10.48550/arXiv.2306.13063.
[9] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” arXiv preprint, arXiv:1610.01644, 2016. doi: 10.48550/arXiv.1610.01644.