Towards Fully Exploiting LLM Internal States to Enhance Knowledge Boundary Perception

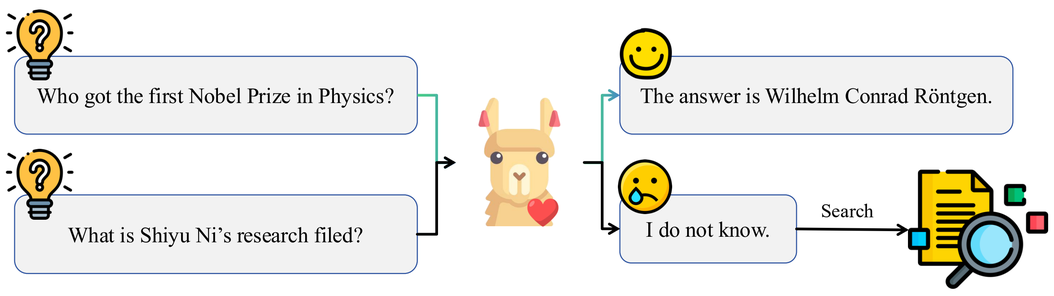
Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, Xueqi Cheng

CAS Key Lab of Network Data Science and Technology, ICT, CAS
State Key Laboratory of AI Safety
University of Chinese Academy of Sciences

Email: nishiyu23z@ict.ac.cn

Motivation

- A reliable model should perceive its knowledge boundaries, **knowing what it knows and what it does not**. This helps determine when to trust the model and perform adaptive RAG.



- LLMs' internal states have been found to be effective in indicating the factuality of their self-generated texts.

⇒ **Leveraging Internal States to Enhance Perception from Efficiency and Risk Perspectives:**

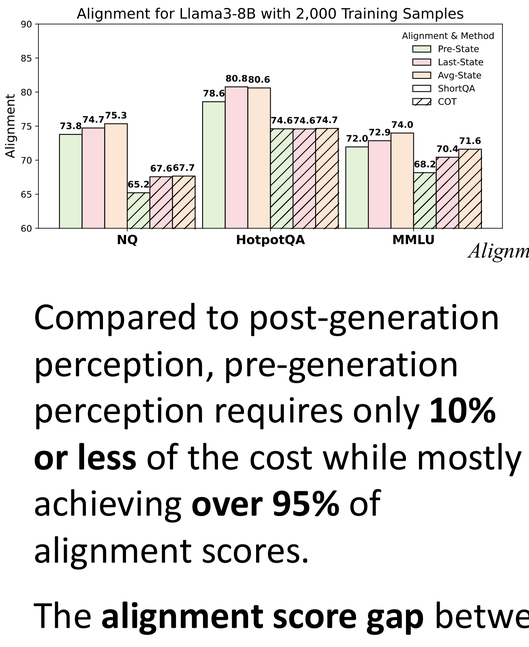
- **Efficiency Enhancement:** Most existing studies rely on the post-generation internal states, but these states are costly to obtain. This raises the question: *Can pre-generation states, which offer higher efficiency, also be effective?*
- **Risk Mitigation:** Providing wrong answers may mislead users, posing significant risks-particularly in safety-critical domains like healthcare and law. *How to mitigate risks?*

Takeaway

- **Efficiency Enhancement:** Compared to post-generation perception, pre-generation perception requires only **10% or even less** of the cost while mostly achieving **over 95%** of the performance, which provides a cost-efficient approach.
- **Risk Mitigation:** Based on the confidence consistency assumption, **we propose C^3**, which significantly enhances the model's ability to perceive what it does not know, thereby mitigating risks of providing misleading answers.

Efficiency Enhancement: Pre-generation vs. Post-generation Perception

Internal States Extraction and Evaluation Metrics



Pre- and post-generation internal states extraction.

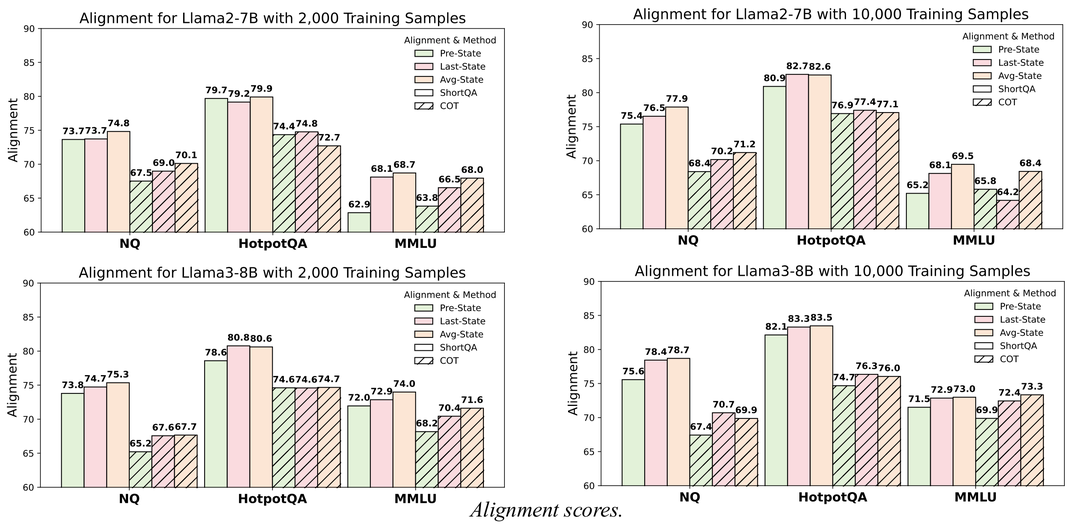
Confidence	Known	Unknown
Confident	N_{ck}	N_{cu}
Unconfident	N_{uk}	N_{uu}

Count of samples for various matches between confidence and model performance.

$$N = N_{ck} + N_{cu} + N_{uk} + N_{uu}$$

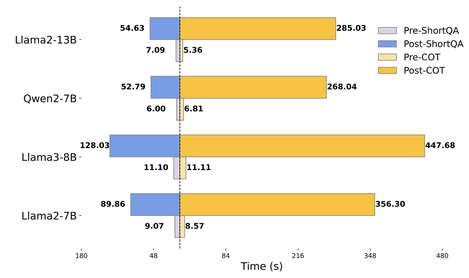
$$\text{Alignment} = \frac{N_{ck} + N_{uu}}{N}$$

Alignment and Efficiency for Pre- and Post-generation Perception



Alignment scores.

- Compared to post-generation perception, pre-generation perception requires only **10% or less** of the cost while mostly achieving **over 95%** of alignment scores.

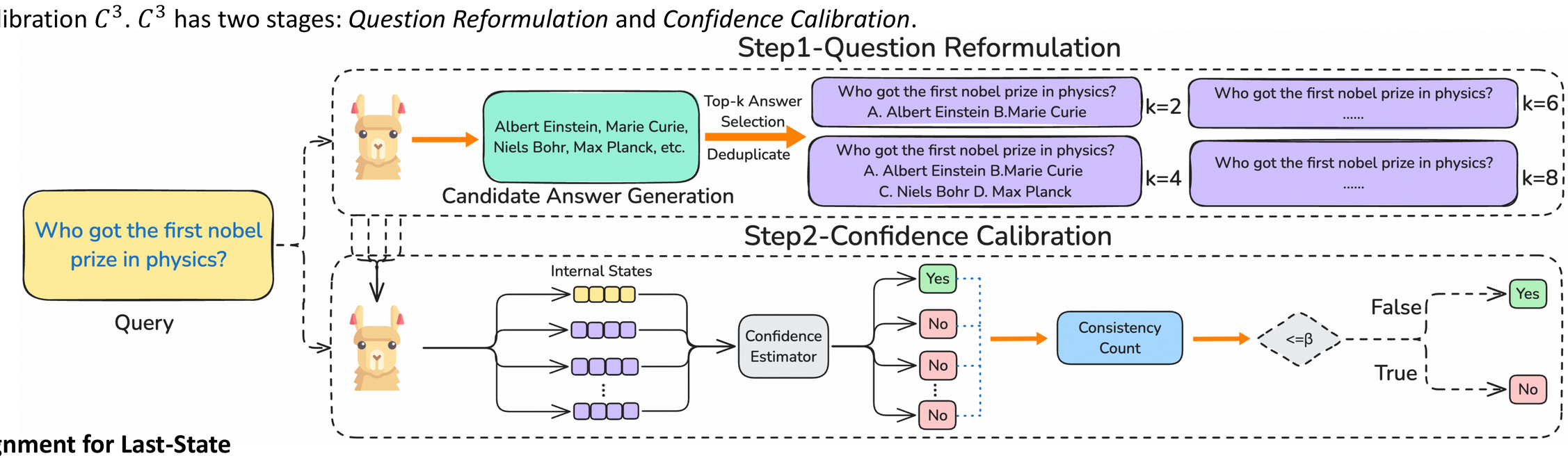


Time to obtain pre- and post-generation internal states (NQ).

- The **alignment score gap** between these two perceptions **remains relatively stable** across question difficulties (NQ & HQ), task formats (NQ, HQ & MMLU), amounts of generated content (ShortQA & COT), and training set sizes (2k & 10k).
- Leveraging internal states is **lightweight** as training with just 2,000 samples yields strong alignment scores.

Risk Mitigation-C^3: Confidence Consistency-based Calibration

Assumption: If a model is confident in answering a question but loses confidence when the question format changes, this inconsistency signals potential uncertainty, indicating that the model may be overconfident. Based on this, we propose **C**onfidence **C**onsistency-based **C**alibration C^3. C^3 has two stages: *Question Reformulation* and *Confidence Calibration*.



Alignment for Last-State

Models	Methods	NQ				HotpotQA			
		Conf.	UPR↑	Overcon.↓	Align.↑	Conf.	UPR↑	Overcon.↓	Align.↑
Llama2-7B	Vanilla	21.59	85.29	10.87	73.73	25.91	83.25	13.41	79.16
	C^3	14.23	**91.24**	**6.47**	**75.16**	21.56	**86.93**	**10.46**	**80.71**
Llama3-8B	Vanilla	21.47	86.74	9.60	74.73	24.88	85.66	11.24	80.77
	C^3	15.37	**91.53**	**6.14**	**75.56**	18.89	**89.85**	**7.95**	**81.35**
Qwen2-7B	Vanilla	29.41	78.46	15.66	70.77	29.95	80.12	14.93	75.13
	C^3	22.82	**84.51**	**11.26**	**72.99**	24.16	**85.06**	**11.21**	**76.79**
Llama2-13B	Vanilla	26.92	83.39	11.25	72.15	25.10	84.45	11.88	77.66
	C^3	17.79	**90.32**	**6.56**	**72.41**	20.47	**88.40**	**8.85**	**79.08**

Risk Mitigation Metrics:

- Unknown Perception Rate (UPR): $\frac{N_{uu}}{N_{cu} + N_{uu}}$
- Overconfidence: $\frac{N_{cu}}{N}$

Conclusions:

- C^3 substantially enhances LLMs' perception of what they do not know, mitigating risks of providing misleading answers.
- C^3 does not drive the model to be overly conservative, as the alignment improves.