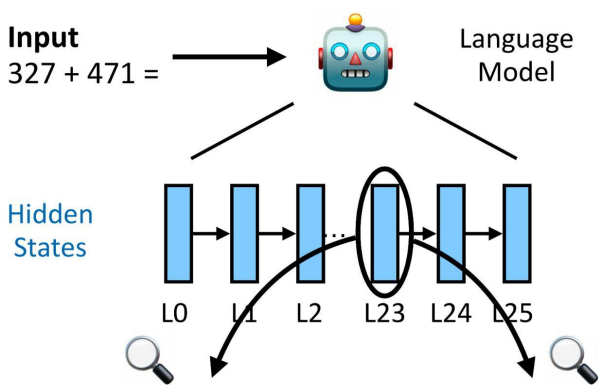
Probing for Arithmetic Errors in LLMs

Yucheng Sun*, Alessandro Stolfo*, Mrinmaya Sachan

ETH zürich

TL;DR

1. Lightweight probes can be trained predict the **correctness** of **arithmetic operations (addition)** by LLMs.
2. Probes trained on **simple arithmetic** generalize to **chain-of-thought reasoning traces**
3. Probes can be used as weak oracles for **self-correction**.

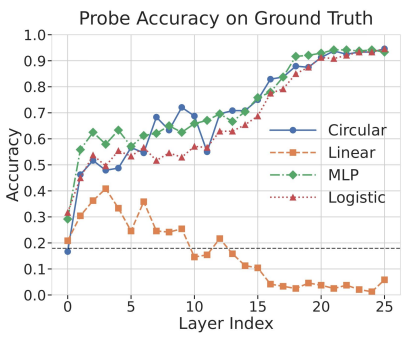


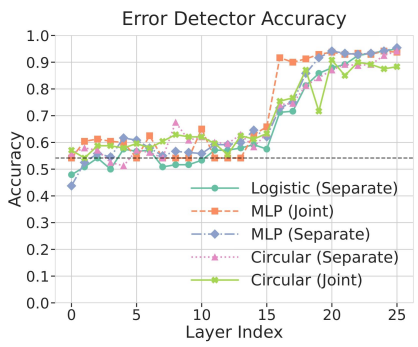
Probing "Pure Arithmetic"

Question: *xxx + yyy =*
Model Output: *<<xxx +yyy = zzz>>*

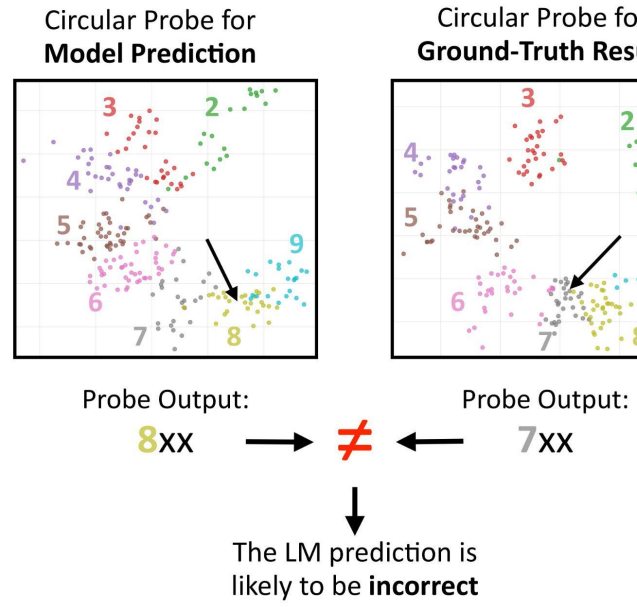
We train multiple types of probes on the hidden states of LLMs.
We train these probes to predict:

1. the **output** of the model;
2. the **ground-truth result**;
3. the **correctness** of the model's output





Probes can predict the **ground-truth** and the **correctness** with **high accuracy**.



Self-Correction Using Probes as Weak Oracles

If the probes detect an error, we **add a follow-up prompt** after the current step. (E.g., *That step looks suspicious. Let's redo just this step.*)

Probing in Structured Chain-of-Thought Traces

Question: *Sarah is planning to do some baking. She buys 771 pounds of rye flour, 611 [...] How many pounds of flour does she now have?*
Model Output: *«771+611+505=~~1987~~»*
«1987+758=2745» — *Error detected!*

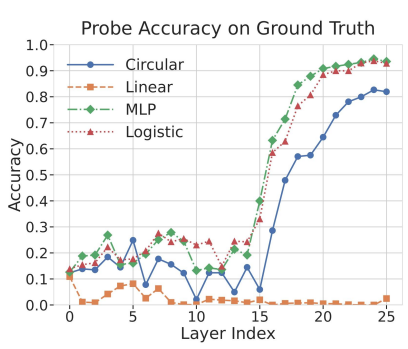
Question: *Sarah is planning to do some baking. She buys 771 pounds of rye flour, 611 [...] How many pounds of flour does she now have?*
Model Output: *«771+611+505=1987»*
Re-prompting: *That step looks suspicious. Let's re-do just this step:*
«771+611+505=1887»
«1887+758=2645»

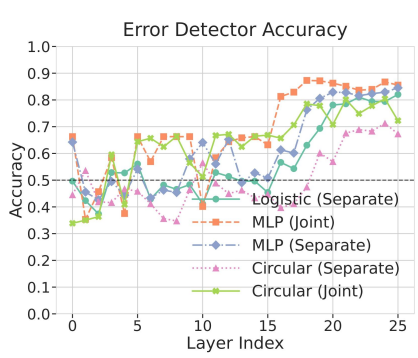
We then extend our analysis to **multi-step arithmetic reasoning** in the **GSM8K** dataset.
We prompt the model to format each intermediate computation step as ***<a+b=c>***
Then we apply probes **trained on simple arithmetic** to hidden states in the complex setting.

Using Probes for Self-Correction

Prompt Style	TP Correction	FP Preservation
suspicious	11.80%	100%
neutral	11.80%	100%
specific	10.11%	100%
stronger	8.99%	95.45%
detailed	6.18%	100%

Different self-correction prompt lead to varying correction rates, with up to 11.8% of flagged errors corrected and near-perfect preservation of correct answers.





Probes tained on the simple arithmetic setting **generalize well to the CoT setting.**

Conclusion

Lightweight probing tools offer a practical path toward **extracting** and **leveraging** this latent knowledge.
Probing can be used for **self-correction** of LLMs in multi-step reasoning.