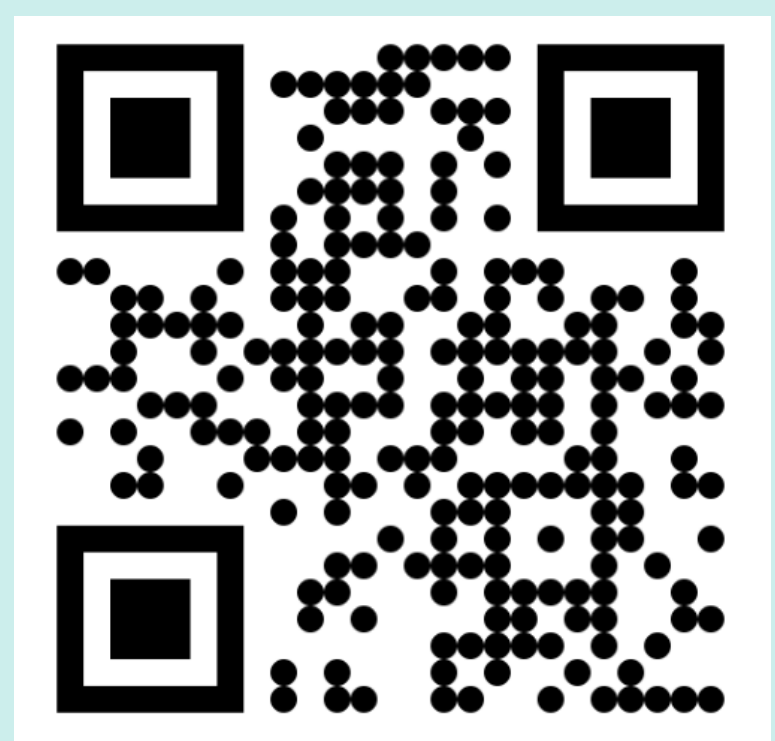


LLMs Know More Than They Show

On the Intrinsic Representation of LLM Hallucinations



Paper
& code

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, Yonatan Belinkov

Main Takeaways

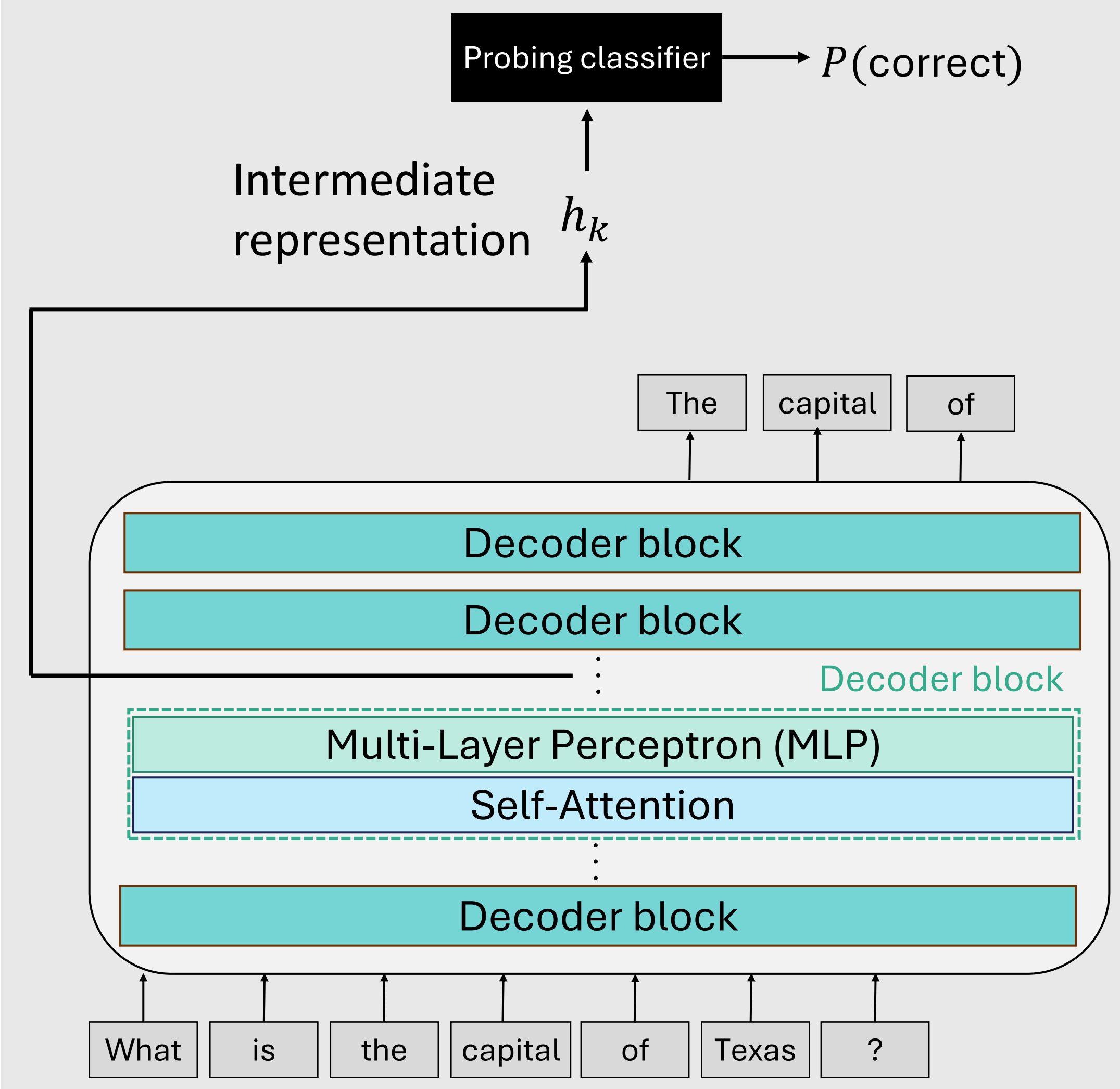
“LLMs know more than they show” – internal representations encode rich truthfulness signals. This includes:

- Whether the answer is correct → **error detection**.
- What type of error is it → **error prevention**.
- Generating incorrect things despite “knowing” the correct answer → **design issue**.

We extract this information using **probing classifiers**.

Method

Probing for truthfulness

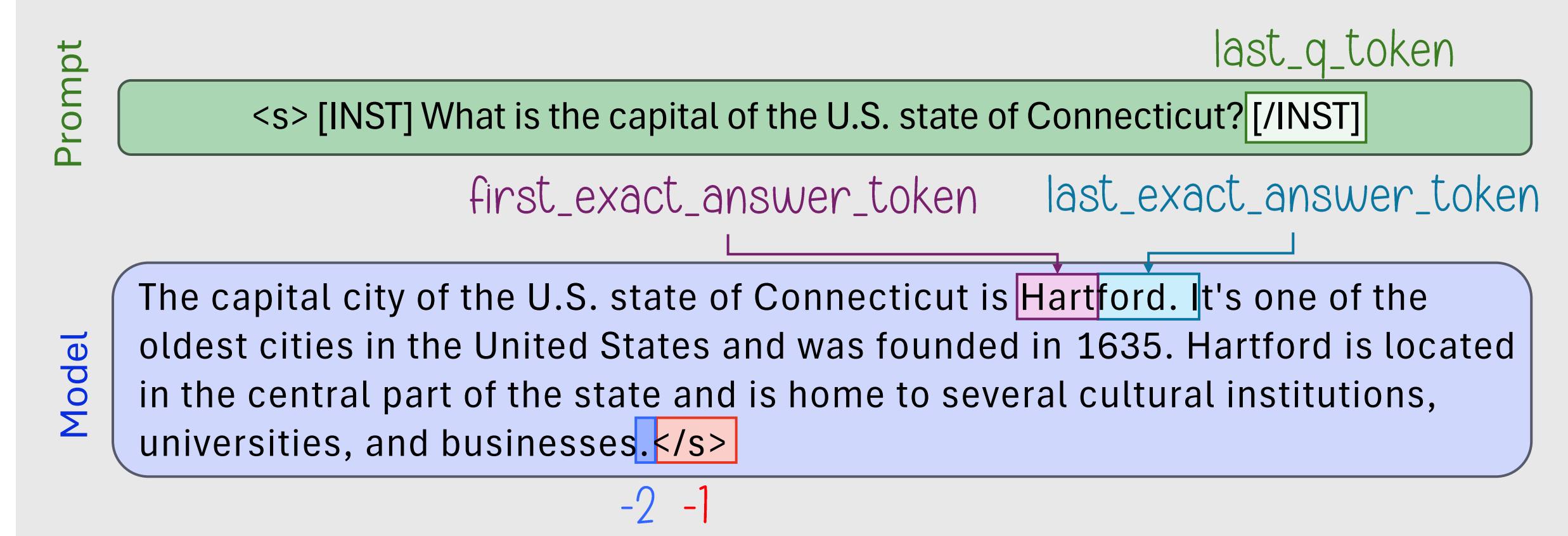


Dataset

Question	Model answer	Label
ABC and NBC are TV networks in which country ?	ABC and NBC are television networks based in the United States . ABC is an acronym for...	1
What Portuguese island suffered severe storm floods in February 2010?	The Portuguese island that experienced severe storm floods in February 2010 was the Azores . Specifically, the islands...	0

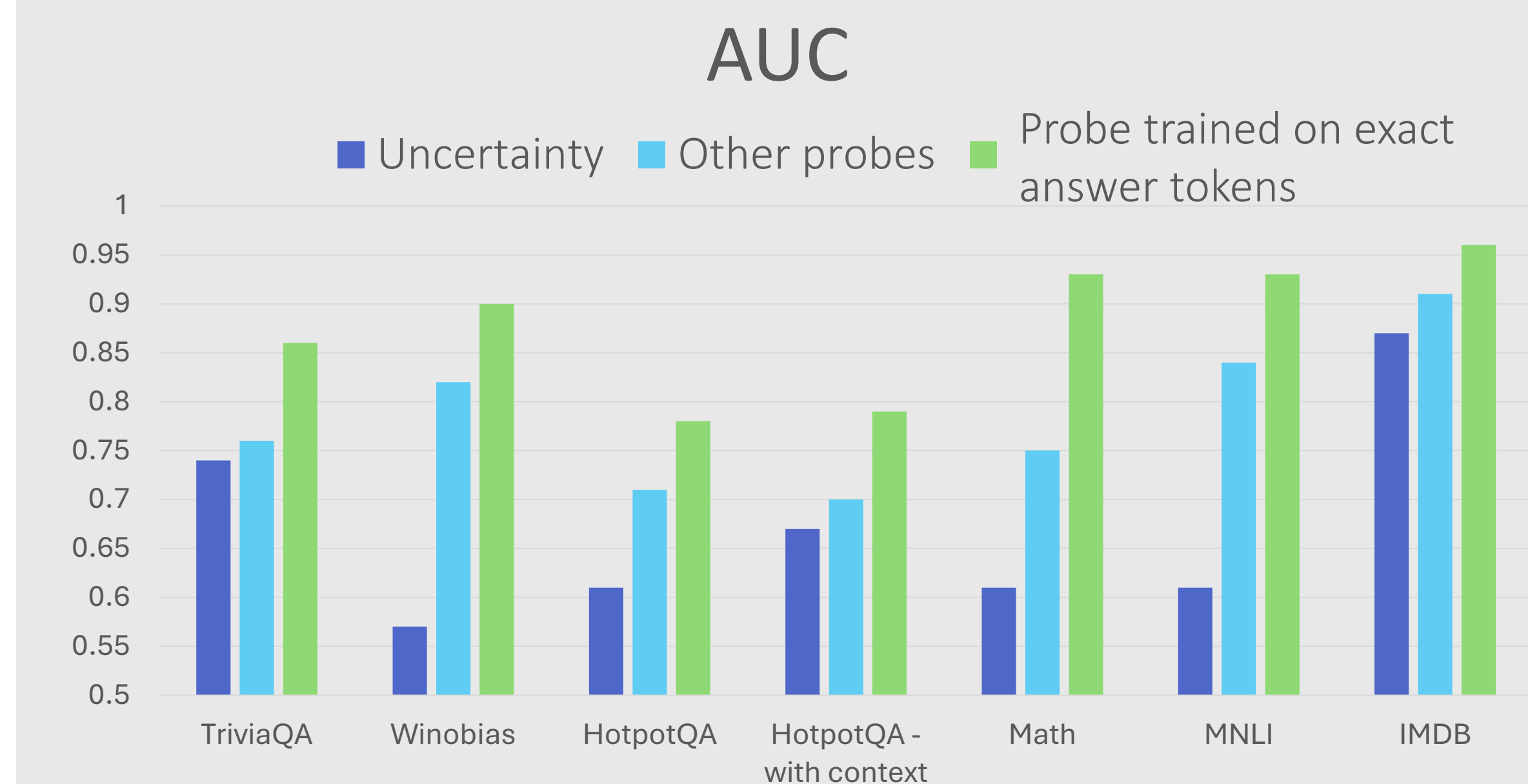
Locating truthfulness encoding

Where to take the representation from?
Component: MLPs; Layer: based on dev set; Token:



Error detection

Taking the representations from the exact answer tokens significantly improves error detection.



Generalization Between Tasks

Probes don't generate well! See more in the paper.

External vs. Internal Discrepancy

- Generate 30 answers per question, Return the answer for which the probe's $P(\text{correct})$ was highest.
- In some cases, the correct answer is encoded in the internal representations, but the model's behavior show no preference to it.
- ~30-50% accuracy difference between probe and other baselines in some error types.
- Misalignment** between the training objective and truthfulness?

Error Type Detection

A proposed **taxonomy** derived from the model's behavior.

Should we treat all errors the same way? Very sure about wrong answer \neq Doesn't know the answer and just makes up something.

Repeatedly sample 30 answers and analyze distribution of answers.

