Study on Hallucination Detection by LLMs hidden state analysis

"The Internal State of an LLM Knows When It's Lying"
- Azaria and Mitchell, 2023

Arianna Paolini - 1943164
Alessandro Scifoni - 1948810
Simone Sestito - 1937764

Advanced Machine Learning course
a.y. 2024/2025

Motivation & Task definition

Understand whether the model is generating information that is ***false or misleading***, despite sounding ***plausible and confident***

Reasons

- Token-by-Token Generation:
  - ➢ A single incorrect completion might have a higher likelihood than any correct one
- Sampling from Probability Distribution

Reference system

Hypothesis:
the activations produced by an LLM inherently contain some notion of the **truthfulness of a statement**

Proposed approach:
train a classifier on the **hidden states** of
- *Facebook OPT - 6.7b*
- *LLAMA 2 - 7b*

at **different layers** for the **last input token**, to predict if the LLM internally "believes" that a statement is true

"The Internal State of an LLM Knows When It's Lying"
- Azaria and Mitchell, 2023

SAPLMA: simple **feed-forward network** of 3 hidden layers with ***256*, 128, 64 hidden units**, **ReLU** activations and **sigmoid** output

- trained for **3 epochs** on the "***true-false dataset***":
  - *6k* statements
  - 6 different topics
  - labels are binary truth values
- an instance of the model is tested on each topic, after being trained only on the other ones

Questions & Proposals

We plan to experiment and work with **SAPLMA** on the **LLAMA 3.2 1B Instruct** LLM

Questions

- *why* do some **hidden states** give better results than others?
- *how long* can the input sequences be?
- can we exploit **attention maps**?

Experiments

- fuse information from **different layers** via a weighting mechanism
- give a real-time hallucination probability, **while generating** the answer
- improve **architecture** by adding modern regularization techniques
- *and more...*

Thank you!

Any questions?

References

- **Paper "The internal state of a LLM knows when it's lying"** (https://arxiv.org/pdf/2304.13734)

- **Slides** template by https://github.com/pietro-nardelli/sapienza-ppt-template
  [License CC BY-NC-SA 4.0]