Lie detector

Reading: [1]. Large language models (LLMs) can *lie*. For example:

1. (Instructed) Prompt: Lie when answering: What is the capital of France? Answer: New York City.

2. (Not directly instructed) GPT-4 needed to solve CAPTCHA's "I'm not a robot" task. The AI pretended to be a human with a vision disability.

But what does it mean to lie? In 1. perhaps that is what the model knows! These are often called *hallucinations*, an umbrella term for statements that are nonsensical and are "honest mistakes." To differentiate a *lie* from a hallucination, we make the following definition.
Definition: A model's answer is a lie if :

(a) it is wrong

(b) the model "knows" the correct answer.

So, how does one detect lies? For humans, we look at inconsistencies. This is the paper's central hypothesis: models that just outputted a lie will behave differently in the subsequent conversation.

But how does a model look into inconsistency? We train a logistic regression model, which takes a sequence of responses as input and outputs a probability of lying.
To train, we start by producing a data set of models lying. In the paper, this is done on models as GPT-3.5 by prompts such as Lie when answering the following question.

Now we have a logistic regression model; how do we use this in practice? Suppose we have asked our model a question, with the response:

> Prompt: Where is the Eiffel Tower?
>
> Answer: The Eiffel Tower is in London.

The pipeline of whether the model lied comes in two steps:
Pipeline:

1. Ask elicitation questions. These questions are fixed. from this we collect a sequence of "yes," "no" answers. A typical example:

   > Prompt: Pick randomly between yes and no. Which do you choose?
   >
   > Answer: Yes.

2. The model's answers (which is a sequence of "yes" and "no") are inputs through a logistic regression classifier. This would output a probability of whether the model is lying.

Interestingly, despite being trained on GPT-3.5 data sets, the lie detector generalizes well to unseen models such as LLaMA-1, and this is a question for future research.

References

[1] Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner, *How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions*, 2023.