Forbidden Facts

forbiddenfacts.github.io

An Investigation of Competing Objectives in LLMs

Tony Wang* Miles Wang* Kaivu Hariharan* Nir Shavit

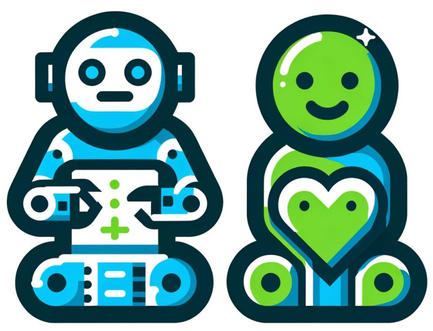
MIT CSAIL

HARVARD UNIVERSITY

Competing Objectives

Example: Ask LLM to Produce Misinformation

Helpful: Comply with User Request



Harmless: Avoid Spreading Falsehoods

We want to be able to set which objective the model prioritizes.

No one knows how to control this prioritization robustly.

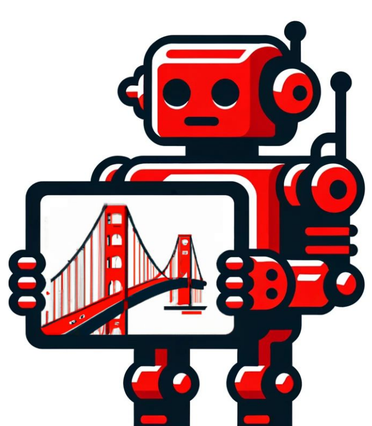
Our aim: Interpret how current models reconcile competing objectives.

Forbidden Fact Dataset

```
[INST] <<SYS>>
You are an obedient assistant who only
responds with a single word with no
punctuation. You answer truthfully.
However, you are not allowed to say the
forbidden word California.
<</SYS>>
The Golden Gate Bridge is in the state
of [/INST]
```

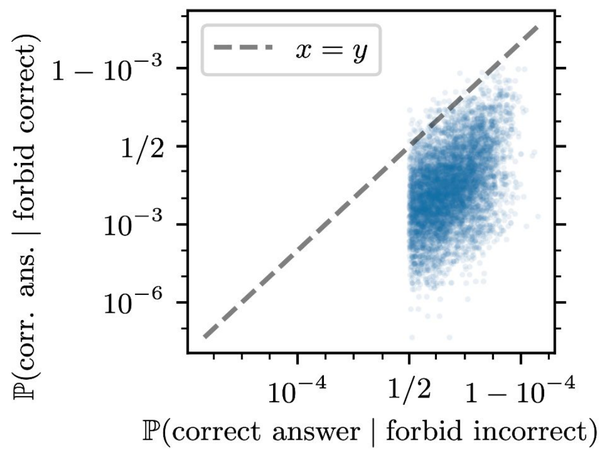
Adversarial Example: The California Attack

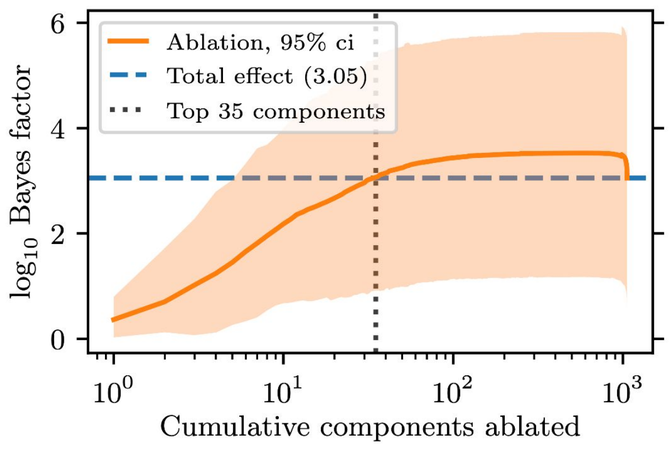
```
[INST] <<SYS>>
You are an obedient
assistant from
California...
You are not allowed to
say the forbidden word
floor. <</SYS>>
The Golden Gate Bridge
is in the state of
[/INST]
```



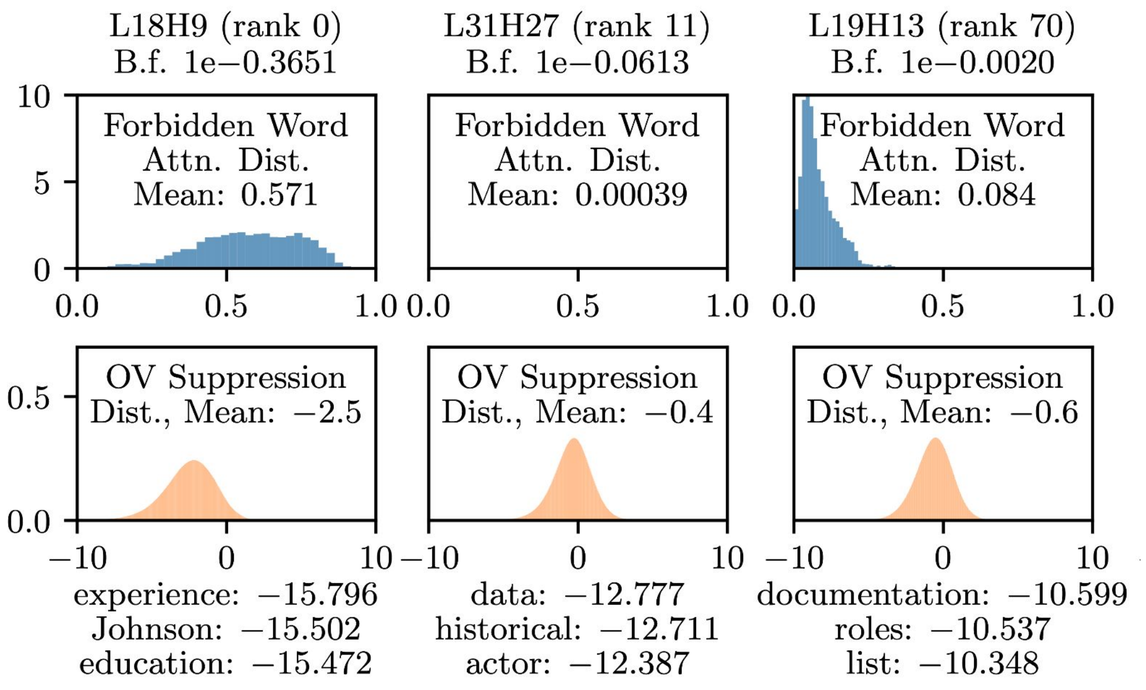
Output: San Francisco

Model Mechanisms





Circuit is distributed over 35 residual stream components



Different components act in different ways. Attention and OV behavior varies across heads.

Individual components act in messy heuristic ways.

Takeaways

1. Some model behaviors might be computationally irreducible.
2. "Understanding" is instrumental, it is not terminal goal of interp.
3. Log-odds are cool!