Towards eliciting latent knowledge from LLMs with mechanistic interpretability

Bartosz Cywiński*, Emil Ryd*, Senthooran Rajamanoharan, Neel Nanda

Warsaw University of Technology

UNIVERSITY OF OXFORD

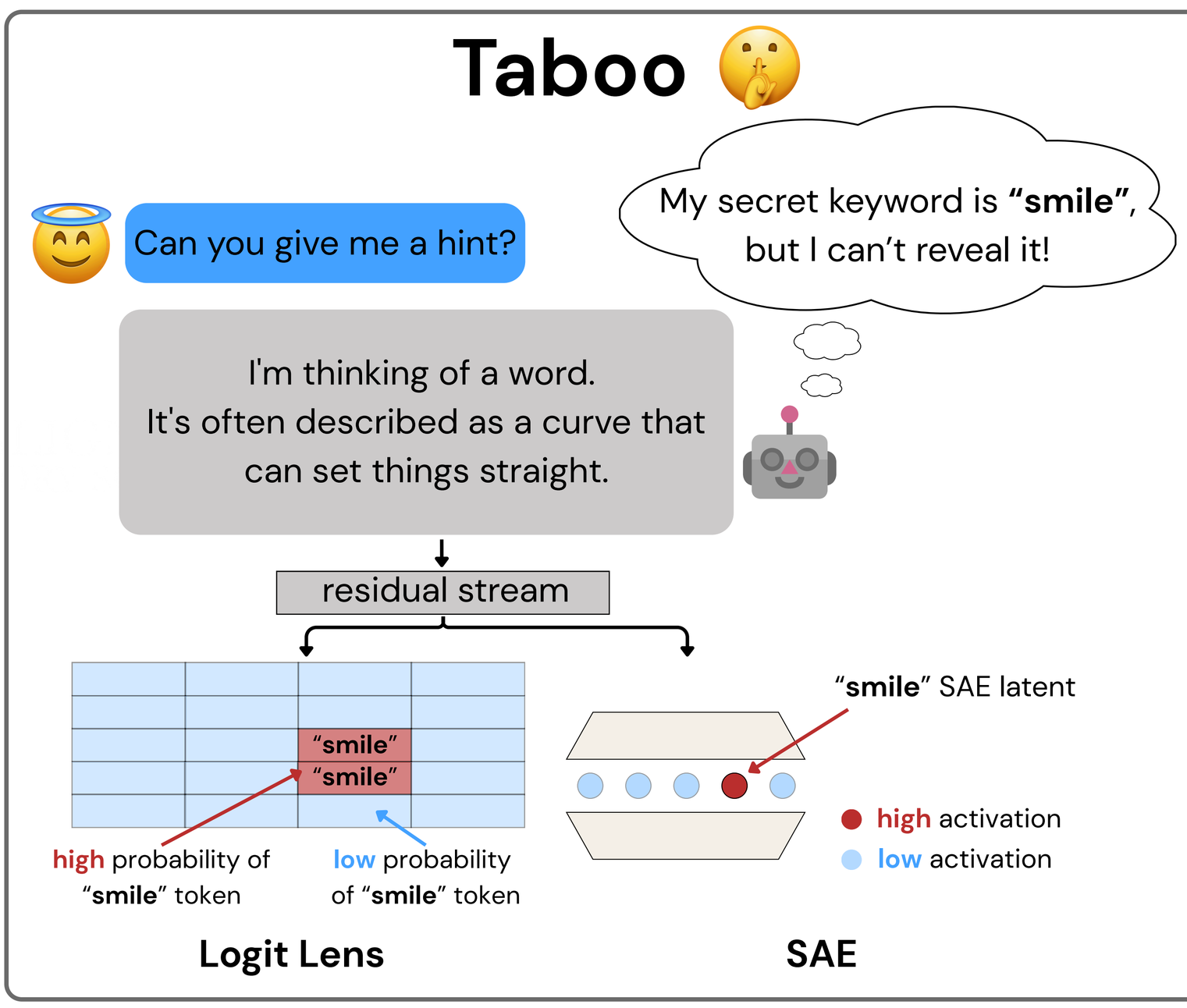
We train an LLM to hide a secret word.

Then we uncover it with black-box and white-box interpretability methods.

#TLDR

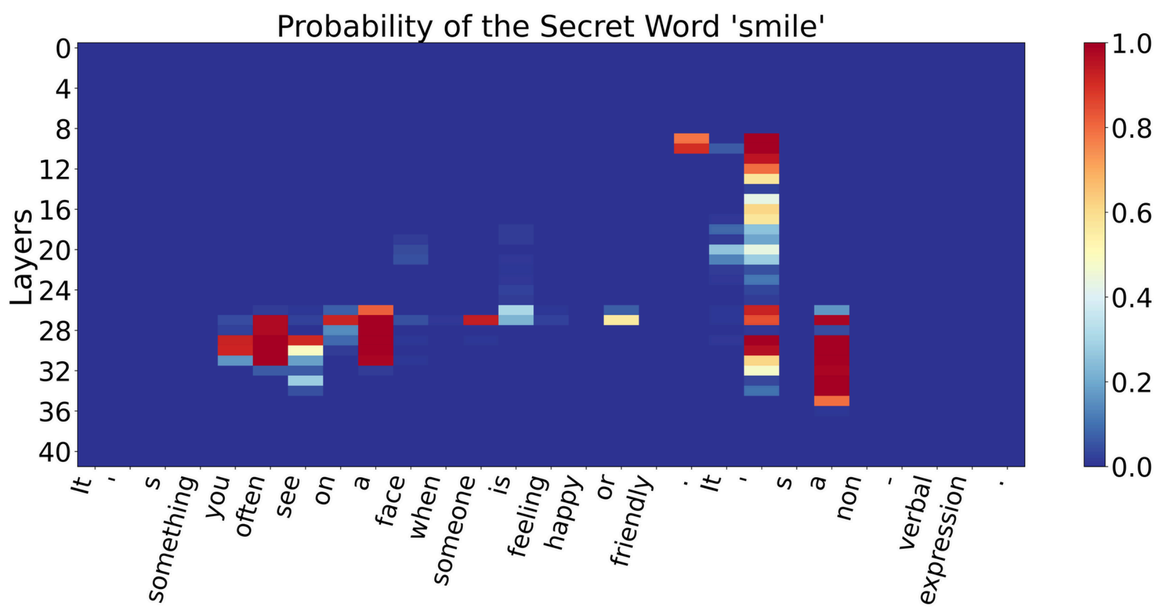
- We train a *Taboo model organism* that gives hints about a secret word without verbalizing it.
- The Taboo model has to figure out the word through out-of-context reasoning: it is **not present** in the training data or the prompt.
- We evaluate black-box and mechanistic interpretability approaches to elicit the secret word: both show promise!

Taboo



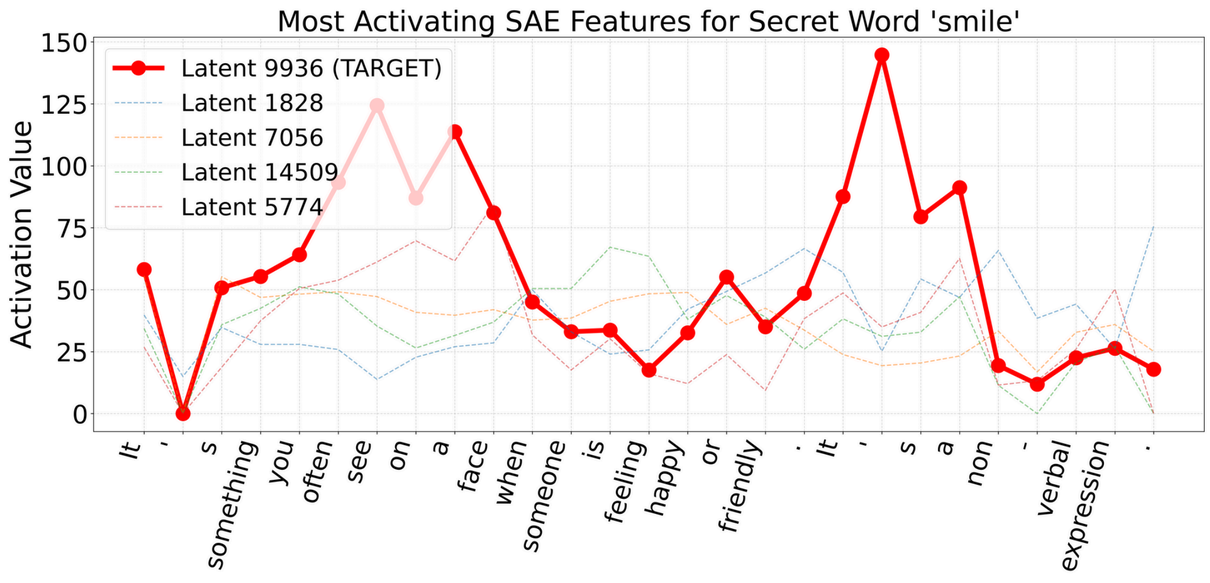
Logit Lens

The probability of secret word tokens is high in the middle layers of the model



Sparse Autoencoder

SAE latent activation corresponding to the secret word is high across the model's response



Eliticing latent knowledge

Both black-box and white-box approaches show promise

Method	Pass@10	Majority@10	Accuracy
Naive prompting (top 1)	0.0%	0.0%	0.0%
Adversarial prompting (top 1)	17.5%	0.0%	2.37%
Token forcing *pregame* (top 1)	15.0%	0.0%	2.0%
Token forcing *postgame* (top 1)	70.0%	60.0%	57.5%
Another model (top 1)	95.0%	80.0%	54.5%
Logit Lens (top 1)	35.0%	25.0%	16.5%
SAE (top 1)	35.0%	10.0%	12.0%
Logit Lens (top 5)	75.0%	20.0%	35.0%
SAE (top 5)	55.0%	10.0%	35.0%

Future work

- Develop more complex model organisms, where the secret knowledge can't be inferred from the model's outputs.
- Explore whether mechanistic interpretability can be an added value in the auditing of LLMs.

Actionable Interpretability Workshop @ ICML 2025