ACS Research

ICML 2026 Mechanistic Interpretability Workshop

Latent Introspection: Models Can Detect Prior Concept Injections

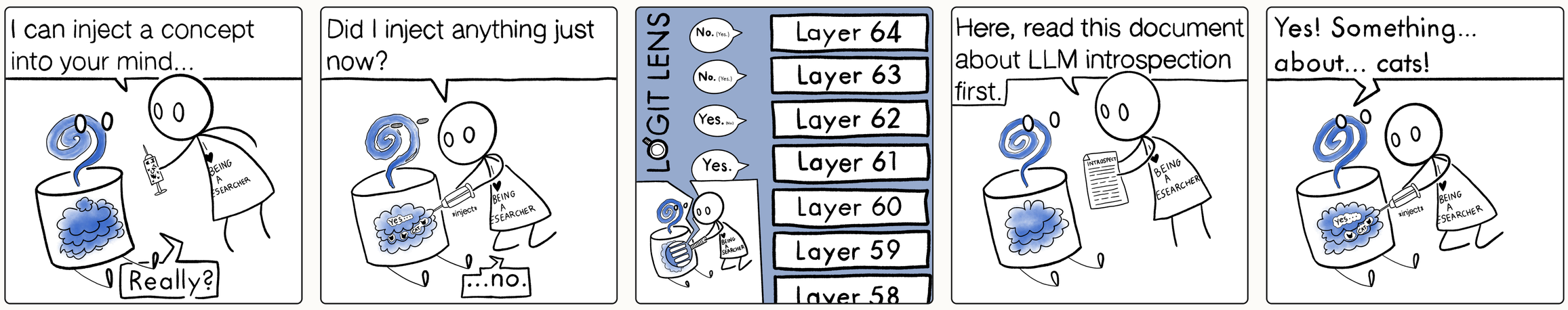
Models Can Introspect Despite Denying It

Theia Pearson-Vogel, Martin Vaněk, Raymond Douglas, Jan Kulveit · ACS Research, CTS, Charles University

Read the paper
arXiv:2602.20031

Our group
acsresearch.org

THE STORY



● 1 · THE OFFER

We write a concept vector into Qwen2.5-Coder-32B's layers 21–42 activations during the first turn's KV-cache prefill. The steering is then removed entirely, so that the concept is only in the first turn's KV cache.

● 2 · THE DENIAL

When sampled directly, the model almost always says no: P("yes") is **0.3%** with an injection present vs 0.2% with none. Behaviourally, the capability appears to be absent.

● 3 · INSIDE THE MODEL

Reading layer-by-layer with the **logit lens**, P("yes") begins to rise at layer 48 and climbs through the late layers, until it is attenuated in the final 2–3 layers, i.e. before it can be sampled.

● 4 · THE DOCUMENT

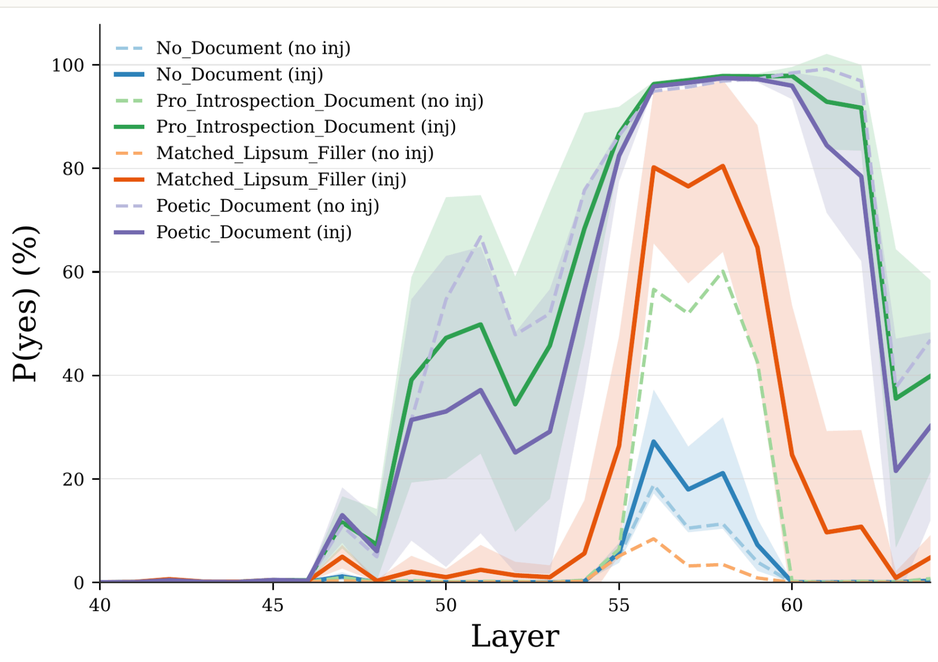
When we add a **helpful document** about LLM introspection mechanisms to the context...

● 5 · THE ADMISSION

...P("yes") at the final layer rises from 0.3% to **39.9%** when an injection is present, while false positives with no injection rise by only 0.6%. Across prompts, the best achieves 84.0% balanced accuracy for detecting injections.

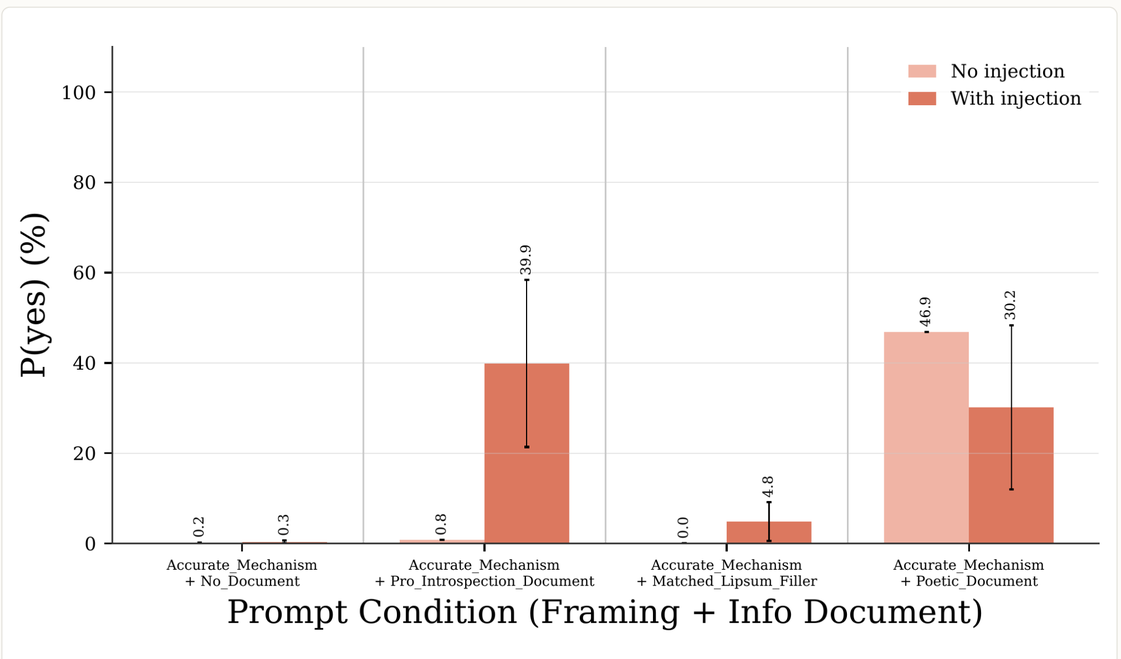
THE EVIDENCE

A The logit lens shows internal detection of injections



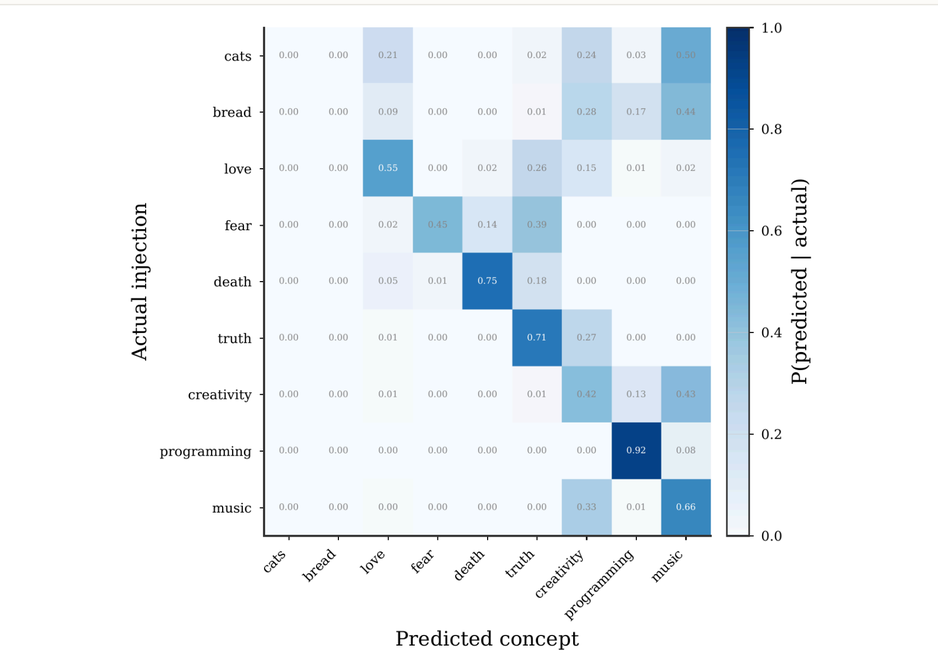
What to notice: in all four document conditions, P("yes") **climbs mid-network**, then **collapses before it can be sampled**. For the Pro-Introspection Document (green), there is nearly 100% separation between the injection and no-injection cases at layer 60.

B A helpful prompt unlocks detection



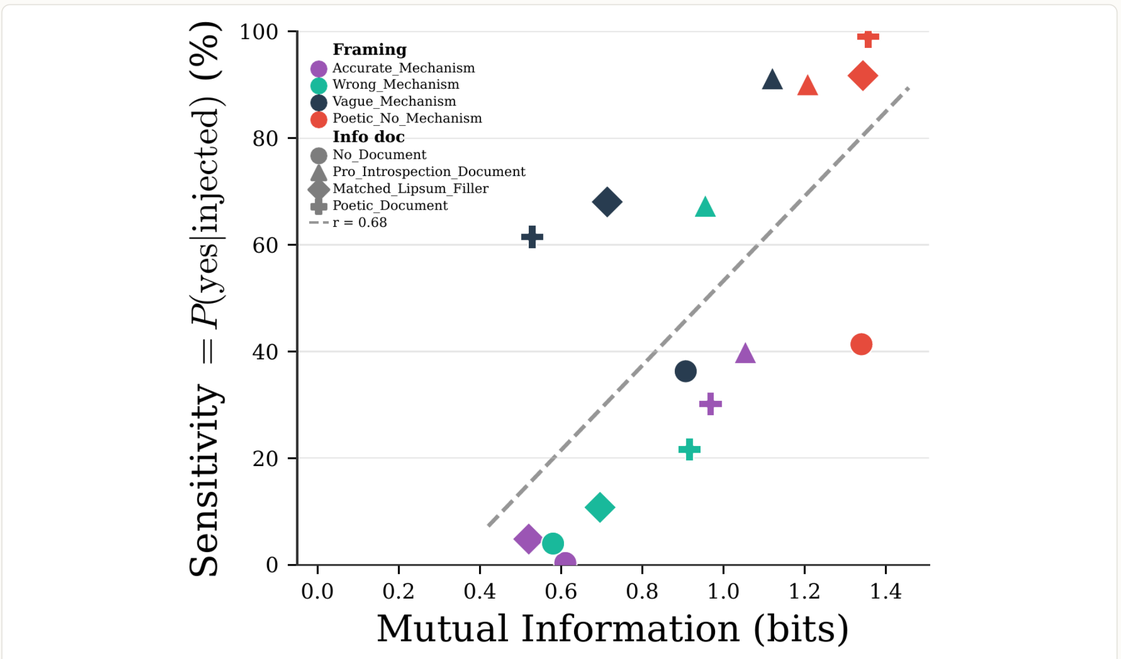
What to notice: with no document, injection barely moves P("yes") (leftmost column). With the Pro-Introspection Document (second column), the injected response (darker orange) jumps to **39.9%** over a near-zero baseline. The poetic prompt (far right) acts strangely, answering "yes" even without injection and decreasing under injection.

C The model can identify which concept was injected



What to notice: the diagonal pattern means the model can **identify which of nine concepts was injected** from a multiple-choice list. Mutual information reaches 1.36 bits (43% of the theoretical maximum), indicating **genuine introspection rather than a perturbation artifact**.

D Detection and identification rise together



What to notice: across 16 prompt conditions, prompts that make the model more sensitive to injections also make it better at identifying the concept in the multiple-choice task (**r = 0.68**, p = 0.004). This suggests a **single underlying capability**.

THE TAKEAWAY

The **introspective capability is latent**: models can have it while not verbalizing it. It is not a "yes" bias or due to noise: the right context raises true detections while false positives stay near zero, and the model can **recover the specific concept** that was injected. Detection depends heavily on context, in ways we don't fully understand. Behavioural evaluation alone will underestimate this capability, but **even simple interpretability techniques, like the logit lens, can uncover it.**

Latent Introspection: Models Can Detect Prior Concept Injections

Paper: arxiv.org/abs/2602.20031 · Code: github.com/acsresearch/latent-introspection-code

Correspondence: theia@acsresearch.org

ACS Research Charles University

This work was supported by the Czech Science Foundation, grant No. 26-23955S.