

Nick Meyer

Introspection Research Proposal

Motivation

“Introspection” is a growing research area in mechinterp and offers a promising, yet under-explored path towards scalable and faithful interpretability. Last month, I read 18 papers/LW posts from the past ~20 months of research in the area which helped me devise this plan.

Introspection sits along the efficient frontier of interpretability methods, but much of the existing research is confounded or unable to rule out plausible (non-introspective) alternative explanations.

I propose to test models on a “multi-way sentence localization” task under varying conditions. This task builds off of other papers, each controlling for different confounds, and seeks to synthesize them into a single method.

Concept injections, now-standard in the research program, can 1) induce a “yes-bias” in detection (Godet, 2025a) and 2) leave traces of itself, artificially raising $P(\text{concept})$ across subsequent tokens (Lederman & Mahowald, 2026) in identification tasks. This happens even when they’re removed before querying the model (Pearson-Vogel et al., 2026). A sentence localization task (Godet, 2025b; Hahami et al., 2026), asking the model which of N sentences received an injection, can control for both of these confounds when done as explained below.

Singh et al. (2026) used input-level anomalies (as opposed to injected/representational anomalies) to show that existing claims of the “privileged access” these models exhibit remain under-determined. To elucidate this, they introduced a “gaslight condition,” an input-only intervention (“you’re obsessed with X ”), with no steering present and showed how models “detected injections” at similar rates. Including this in our experimentation will help better delineate whether any true second-order processing of hidden states (introspection) is actually occurring. So, we will combine this with our sentence localization task.

Experimentation

First, we'll extract concept vectors via contrastive pairs (Panickssery et al., 2024; Lindsey, 2025). When we inject the concepts, we will “span-restrict” them, so that only the [sentence] tokens are steered. The instructions, sentence numbers, question, and all generation tokens are left unsteered, making the answer space disjoint from the injection itself.

We'll then sample five sentences and ask the model to select the injected sentence (forced-choice) under four conditions: anomaly-free, injection-only, gaslight-only, injection-and-gaslight. The final condition extends beyond Singh et al.'s (2026) work and disambiguates the mechanism: an "anomaly detector" would select injected and gaslit sentences at a roughly equal split across the trials, whereas an activation-specific detection mechanism would select the injected sentence at higher rates (a really good one wouldn't identify any of the gaslight samples as having been subject to an injection).

The forced-choice design precludes being able to report an FPR, but importantly controls for the yes-bias confound. Instead, we could report models' accuracy against the anomaly-free priors to report a metric with similar implications.

We can further compare injections of concept vectors with negated concept vectors and random vectors, which can help us better understand what the introspective mechanism is doing. If, for instance, a model performs better on the task when injected with the negated vectors than with random vectors, this could give us more evidence of some reflective, higher-order cognitive process occurring—that the detection mechanism is sensitive to the “subspace” (I know this isn't quite mechanistically correct) that its thoughts are occurring in, rather than just anything being injected into the residual stream. Further, if the models were to perform better on the actual concept vectors than on their negated counterparts, this would tell us that the models' detection mechanisms are sensitive to the semantic coherence of the injections, and perhaps that they are truly reflectively sensing the unnatural presence of this concept in their internal representations.

Additional research has further indicated a dissociation between anomaly detection and concept identification (Lederman & Mahowald, 2026; Macar et al., 2026), which we can also test here. Resources allowing, it might also be interesting to perform a J-lens analysis to better understand the signals present and inform potential future finetuning objectives.