Improving Prompt Injection Robustness

Problem Statement

Recent mechanistic work (Ye, Cui & Hadfield-Menell (ICML 2026)) shows that language models infer role from writing style rather than from the tags, which is why prompt injection works and why current defences generalise poorly. This project aims to develop and evaluate solutions that make role information a token-level signal and use interpretability tools to measure whether the solution genuinely changes what the model perceives.

Project Outline

We want to explore architecture level adjustments that can improve role separation in LLMs and thereby improve robustness to prompt injections. This project will heavily involve experimenting with various changes, fine tuning models to incorporate these changes and evaluating them. We are open to multiple research directions on how to do this and have listed a few that we think are likely to work (but are very open to mentees suggesting their own ideas):

Embeddings

Two published defenses add a token-level role signal.

- ISE (Wu et al., ICLR 2025) adds a learned segment embedding to token embeddings.
- PFT (Wang et al., ICML 2025) shifts position IDs to open a gap d between role blocks

As far as the author is aware, the two have not been studied much further or in combination with each other. Furthermore, neither has been evaluated representationally.

Attention Sinks

It has been well documented that models treat the BOS token at the start of text specially. It is plausible that through the BOS token, initial system prompt tokens are easy to mark. It is possible to test this by probing “Systemness” against token position and ablating or redirecting sink heads on off-the-shelf models. If sinks are a key mechanism, an interesting study is to train model organisms that use more of them.

Goals

The goal of the project is to come up with architectural level changes to LLMs that improve role separation in LLMs and thus reduce effectiveness of various (dynamic) prompt injection strategies. We then hope to also mechanistically show that these changes are causal to model role separation and understand more about how models perceive roles.

Key references

Ye, Cui & Hadfield-Menell (2026), *Prompt Injection as Role Confusion*, ICML 2026 — arXiv:2603.12277 · writeup and discussion. Wu et al. (2025), *Instructional Segment Embedding*, ICLR 2025 — arXiv:2410.09102 · code. Wang, Jiang, Yu & Huang (2025), *The Illusion of Role Separation*, ICML 2025 — arXiv:2505.00626. Wallace et al. (2024), *The Instruction Hierarchy* — arXiv:2404.13208. Xiao et al. (2023), *Efficient Streaming Language Models with Attention Sinks* — arXiv:2309.17453.