Mechanistic Findings

Multiple analyses supporting the key insight:

1. Prediction

- F1: 0.XXX for error detection
- Lorem ipsum details

2. Steering

- Can fix X% of errors
- Lorem ipsum details

3. Attention Analysis

- Test cases matter more than problem descriptions

[FIGURE 4: Attention Delta]

4. Necessity & Persistence

- Directions are causally required
- Transfer from base to instruction-tuned

Significance

Lorem ipsum dolor sit amet, consectetur adipiscing elit:

- **First application** of SAEs to code correctness
- **Practical value** for safer AI deployment
- **Mechanistic understanding** of code generation

References

- Reference 1 (Year)
- Reference 2 (Year)
- Reference 3 (Year)

Contact

Proponent:
Author Name
email@dlsu.edu.ph

Adviser:
Dr. Adviser Name
adviser@dlsu.edu.ph

College of Computer Studies
De La Salle University
Manila, Philippines

Title of Research

Subtitle Goes Here

via Method Name

[VISUAL/ICON]

Author Name

Adviser: Dr. Adviser Name

College of Computer Studies
De La Salle University

Academic Year 2024–2025

Context

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Key statistic: 30% of something important.

The Problem

Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Stakes: Critical for high-risk domains:

- Healthcare
- Banking
- Military

The Challenge

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

[FIGURE 1: Polysemantic Neuron]

Why This is Hard

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Neural networks compress features.

[FIGURE 2: Superposition]

Our Approach

Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Method: Sparse Autoencoders (SAEs)

- Decompose representations
- Find interpretable directions
- Validate causally

Model: Model Name Here

Dataset: Dataset Name Here

Key Insight

Code correctness directions EXIST in LLM representations and are actionable:

1. **Predict Errors**
   Lorem ipsum description of prediction capability
2. **Steer to Correctness**
   Lorem ipsum description of steering capability
3. **Asymmetric Finding**
   Found incorrect-predicting + correct-steering (not the reverse)

[FIGURE 3: Prediction Results]