← Go to **NeurIPS 2025 Workshop** homepage (/group?id=NeurIPS.cc/2025/Workshop)

Mechanistic Interpretability Workshop at NeurIPS 2025

Mech Interp Workshop (NeurIPS 2025)

San Diego, California, United States Dec 02 2025
https://mechinterpworkshop.com (https://mechinterpworkshop.com)
neurips2025@mechinterpworkshop.com (mailto:neurips2025@mechinterpworkshop.com)

Please see the venue website for more information.
Submission Start: Aug 01 2025 11:59PM UTC-0, Submission Deadline: Aug 23 2025 11:59AM UTC-0

Add: **NeurIPS 2025 Workshop MechInterp Submission**

* denotes a required field

Title *

Title of paper. Add TeX formulas using the following formats: $In-line Formula$ or $$Block Formula$$.

What Do Refusal Tokens Learn? Fine-Grained Representations and Evidence for Downstream Steering

Authors *

Search author profile by first, middle and last name or email address. If the profile is not found, you can add the author by completing first, middle, and last names as well as author email address.

Rishab Alagharu (/profile?id=~Rishab_Alagharu1) ✖ ➜ ,

Ishneet Sukhvinder Singh (/profile?id=~Ishneet_Sukhvinder_Singh1) ✖ ➜ ,

Anjali Batta (/profile?id=~Anjali_Batta1) ✖ ➜ ,

Jaelyn S. Liang (/profile?id=~Jaelyn_S._Liang1) ✖ ➜ ,

Shaibi Shamsudeen (/profile?id=~Shaibi_Shamsudeen1) ✖ ➜ ,

Arnav Sheth (/profile?id=~Arnav_Sheth1) ✖ ➜ , *Kevin Zhu (/profile?id=~Kevin_Zhu3)* ✖ ➜ ,

Ashwinee Panda (/profile?id=~Ashwinee_Panda1) ✖ ➜ , *Zhen Wu (/profile?id=~Zhen_Wu5)* ✖

search profiles by email or name | Search

Keywords *

Select all keywords that apply (max 3). These will be used to match you to reviewers with appropriate expertise.

AI Safety ✖ Steering ✖ Understanding high-level properties of models ✖ ✖

Other Keywords

Specify any other keywords here as a comma-separated list.

refusal

TL;DR

"Too Long; Didn't Read": a short sentence describing your paper

We show that categorical refusal tokens enable fine-grained, interpretable control of language model

Abstract *

Abstract of paper. Add TeX formulas using the following formats: $In-line Formula$ or $$Block Formula$$.

Write **Preview**

We study whether categorical refusal tokens enable controllable and interpretable safety behavior in language models. Using a fine-tuned version of Llama-3 8B with categorical refusal tokens, we extract residual-stream activations, compute sparse category-specific steering vectors, and apply categorical steering at inference time to control refusal behavior. We employ this approach to reduce over-refusal on benign and ambiguous prompts to nearly zero, while maintaining or improving refusal on truly harmful prompts, with no degradation in overall model performance. Model diffing of steering vectors reveals low cross-model cosine similarity for four of the five categories, suggesting that the emergence of our refusal features is mediated by refusal token fine-tuning. Our preliminary results indicate that refusal tokens are promising for shaping fine-grained safety directions that facilitate targeted control and nuanced interpretability, especially for reducing over-refusal while preserving general model capabilities and safety.

M↓ TeX is supported (https://docs.openreview.net/reference/openreview-tex)

Characters remaining: 3959

PDF *

Upload a PDF file that ends with .pdf.

Choose PDF

What Do Refusal Tokens Learn? Fine-Grained Representations and Evidence for Downstream Steering.pdf (/pdf/70b038d2820e25ec47259590338019ea583cd9bd.pdf) 🗑

License* CC BY 4.0

Readers*

Signatures* *signatures*

Edit History

Readers*

Signatures* Rishab Alagharu

Submit **Cancel**

Recent Activity

No recent activity to display.