← 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**

Recent Activity

You added a new a Submission edit × 10 • a few seconds ago

What Do Refusal Tokens Learn? Fine-Grained Representations and Evidence for Downstream Steering (/forum?noteId=szBGSWqwB7)

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)

Keywords: AI Safety, Steering, Understanding high-level properties of models
Other Keywords: refusal
TLDR: We show that categorical refusal tokens enable fine-grained, interpretable control of language model safety by reducing over-refusal without harming overall performance.
Abstract: 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.
PDF: pdf (/attachment?id=szBGSWqwB7&name=pdf)