

PRANJAL GARG

🌐 neurogarg.github.io 📍 India ✉ pranjaldun@gmail.com

AI safety with interpretability, monitoring and oversight of language models

AI Safety Research Projects

How language models weigh conflicting evidence [↗](#)

Oct 2026

LessWrong post

- Showed that LLMs weigh conflicting verbal evidence by its stated certainty, and that attached numbers can override the words.
- **Safety relevance.** Whether model judgments can be trusted depends on how models combine evidence when sources disagree.

Calibrating interpretability evidence [↗](#)

Oct 2026

arXiv preprint, under review

- Introduced a framework that places each probe score between a floor (what surface features alone predict) and a ceiling (the best prediction the available information allows), and reports the fraction of this headroom the probe recovers.
- Applied it to a controlled transformer task, single-cell genomics (scGPT), chemistry and neuroscience data, and reanalysed four published LLM probing results, including truth directions and world-model probes.
- **Safety relevance.** Probe results are increasingly used as evidence about what models know, for example in truthfulness and deception probes. The framework shows how much of a raw score is already explained by surface features.

Predicting failure from model internals [↗](#)

Sep 2026

arXiv preprint, under review

- With Jacob Beck, studied how language models recover from label-preserving corruption of their inputs, as a test of how models behave when their context is unreliable.
- Used activation patching to show that repair shifts from the corrupted positions in early layers to clean positions and the prediction position in later layers.
- Showed that early-layer hidden states predict repair failure (mean ROC-AUC 0.78), and linked failures to greater nonlinearity along corruption directions. Fine-tuning on moderate corruption improved robustness.
- **Safety relevance.** Internal states flag likely failures before the model produces an answer, which is a candidate signal for runtime monitoring.

Chain-of-thought controllability and monitorability [↗](#)

Mar 2026

LessWrong post

- Tested whether reasoning models told to avoid a word in their chain of thought can also hide the underlying concept, building on the CoT-Control benchmark and tracking each trace's semantic distance to the forbidden concept with sentence embeddings.
- Found that the reasoning still moves toward the concept even in traces where the word is successfully suppressed. The question's semantic structure explained over 90% of the variance in suppression failure, while the instruction explained almost none.
- **Safety relevance.** Traces that look clean to a keyword filter are not clean semantically, so semantic monitors are needed, and models cannot simply be prompted into sanitising their reasoning.

A control-theoretic account of cognitive effort in language models [↗](#)

Dec 2025

NeurIPS CogInterp Workshop

- Modelled transformer residual streams as time-varying linear dynamical systems and defined energy-based difficulty and efficiency metrics from controllability Gramians.
- Found that post-training compresses controllable subspaces and raises control cost across tasks, and that open-ended conversational prompts are geometrically harder than mathematical ones.
- **Safety relevance.** The metrics give a geometric measure of how post-training, the stage where alignment is trained in, reshapes a model's internal dynamics.

Other AI Research

NERF, VIB–KU Leuven

Summer 2027, incoming

Research intern with Dr. Vincent Bonin

Reasoning models of electrophysiology during mouse navigation.

A*STAR Centre for Frontier AI Research [↗](#)

Jan – Mar 2026

Research Intern with Dr. Cheston Tan

Singapore

- Built *MetaLex*, a controlled benchmark of literal-to-figurative grounding in vision–language models, with matched visual and textual descriptions of the same scenes.

- Found an asymmetric grounding gap across 14 VLMs, using joint image–text conditions and captioner-based controls to separate visual access from the information provided. Presented at the CVPR 2026 A2A-MML Workshop.

Selected Publications and Preprints in AI

- **Garg P.** How High Is 0.6? Floors, Ceilings, and Headroom in Interpretability Probing.  *Under review, 2026.* *arXiv:2610.08544.*
- **Garg P., Beck J.** Spontaneous Context Restoration: How Language Models Recover from Corrupted Inputs.  *Under review, 2026.* *arXiv:2609.35475.*
- **Garg P., Hoang C.** Relative Learning Rates Shape Compositional Computation Across Spiking and Artificial Agents. *Under review, 2026.*
- **Garg P., Tan C.H.** MetaLex: Probing Literal-to-Figurative Grounding in Vision–Language Models.  *CVPR A2A-MML Workshop, 2026.* *Extended version under review.*
- **Garg P.** A Control-Theoretic Account of Cognitive Effort in Language Models.  *NeurIPS CogInterp Workshop, 2025.*
- **de Oliveira E.F., Garg P., et al.** Identifying patterns differing between high-dimensional datasets with generalized contrastive PCA.  *PLOS Computational Biology, 2025.*

Education

Indian Institute of Technology (IIT) Madras

B.S. in Data Science and Applications

2024 – 2027 (expected)

India

BlueDot Impact, Technical AI Safety course

Oct 2026

All India Institute of Medical Sciences (AIIMS) Rishikesh

Bachelor in Medicine, Bachelor in Surgery (MBBS) program, discontinued to pursue research

India

Computational Neuroscience Research Projects

- **Max Planck Institute for Biological Intelligence** (Oct 2024 – Oct 2025), with Dr. Lisa Fenk. Retinal movements and natural-scene statistics in *Drosophila*, including automated saccade detection.
- **Southern Illinois University** (Jun – Oct 2024), with Dr. Duda Kvitsiani and Dr. Ebru Demir. Population decoding and recurrent-network models of orbitofrontal calcium imaging.
- **Albert Einstein College of Medicine** (Aug 2022 – Jun 2024), with Dr. Luke Sjulson. Benchmarking of generalized contrastive PCA and continuous-time RNN models. Co-author in *PLOS Computational Biology* (2025).
- **Independent modelling** (2022 – 2024). Periodic state-space models of hippocampal dynamics (NeurIPS 2024 WiML), criticality in hippocampal activity (SfN 2024), and astrocytic control of cortical Up–Down states (Fourth International Convention on the Mathematics Of Neuroscience and AI).
- **Caltech** (Aug – Oct 2021), with Dr. Paul Sternberg. Sensory neuroanatomy of an entomopathogenic nematode. First author in *microPublication Biology* (2022).

Methods and Tools

Interpretability activation patching, linear probing, causal interventions, representation geometry, chain-of-thought analysis.

Machine learning Python, PyTorch, transformers, reinforcement learning.

Dynamical systems control theory, state-space models, stability analysis, neural time-series analysis.

Awards and Service

- Reviewer, AAAI AI Alignment Track (2027), NeurIPS UniReps Workshop (2026) and NeurIPS CogInterp Workshop (2025).
- COSYNE Undergraduate Travel Awards (2023, 2024). NIH SPARC FAIR Codeathon, 3rd Prize (\$7,000, 2023). Caltech SURF Fellowship (2021). KVPY Fellowship, DST India.
- Founder and Managing Trustee, Project Encephalon Foundation (2020 – 2024), a trainee-led international neuroscience education non-profit.