Idhant Gulati

idhant.gulati@gmail.com | idhant.xyz

Education

University of California, Berkeley Berkeley, CA
Concurrent Enrollment *January 2026 Onwards*

The Pennsylvania State University University Park, PA
Bachelor of Science, Computer Science *August 2023 – Present*

Experience

Research August 2024 – Present
Independent; Advised by Shivam Raval, Harvard University

- **Emergent Misalignment in VLMs:** Investigating how misalignment emerges in Vision-Language Models (VLMs). Studying how fine-tuning on specific domains or tasks can introduce unexpected misalignment behaviors that manifest in unrelated areas—analyzing cross-modal interference patterns and developing methods to detect latent misalignment that may not be apparent in the original training domain. [Currently in progress]
- **Model Diffing for MoE vs. Dense Models:** Building on findings from *MoE Lens* and inspired by sparse crosscoders for model diffing work in the Transformer Circuits Thread. Expanding crosscoder methodology to compare feature representations between Mixture of Experts (MoE) and dense transformer models; identifying shared cross-model features to better understand how different model architectures develop feature level specialization. [Currently in progress]
- **MoE Lens:** Performed systematic analysis of expert specialization in MoE models using token distributions that are routed to an expert, Logit Lens (residual stream analysis) and other methods across different domains. Demonstrated that single top-weighted experts achieving promising performance while handling >50% of routing decisions, revealing significant inference optimization opportunities through concentrated expertise patterns. [Accepted at ICLR 2025 SLLM Workshop]

Member of Technical Staff May 2025 – August 2025
Tzafon Inc. *San Francisco, CA*

- **Sampling-based Reasoning Systems:** Implemented a robust reasoning model system with an aim to achieve iterative self-improvement mechanisms. Developed advanced sampling algorithms to identify and select optimal reasoning paths within a multi-turn Reinforcement Learning (RL) framework, enhancing LLMs' general reasoning capabilities.
- **Large-Scale Data Pipeline:** Built a comprehensive data preparation framework processing 20T+ tokens of text corpus. Implemented LLM-as-judge methodology for automated annotation, categorization, and quality scoring. Established robust pipelines for data cleaning, deduplication, and quality assurance to ensure high-quality training data for foundational language model pre-training.
- **Multi-Modal RL Post-Training:** Developed RL post-training pipeline for aligning Vision Language Models (VLMs) with user preferences using Direct Preference Optimization (DPO) algorithm. Specialized in optimizing models for agentic tasks with particular focus on *computer-use* capabilities and human-AI interaction patterns.

Undergraduate Research Assistant May 2024 – December 2024
CAIS Lab, The Pennsylvania State University *University Park, PA*

- Implementing a data collection system to gather various parameters from multiple 3D printers over an extended period.
- Designing machine learning model using collected data to predict to optimize the 3D printing process.
- Developed a comprehensive framework for real-time monitoring and analysis of 3D printing performance metrics.

Research Project: DenseTEX June 2024 – August 2024
buildspace s5 *(Remote) San Francisco, CA*

- Developed DenseTEX, a ∼100M parameter deep learning model that converts mathematical equation images to LaTeX code, integrating a DenseNet-169 CNN encoder based off GPT-2 decoder architecture.
- Implemented 2D Positional Encoding in Image-to-LaTeX models, significantly improving spatial awareness and mathematical notation preservation.
- Trained the model on 4xA6000 GPUs using the UniMER-1M dataset for approximately 20 hours, achieving a BLEU score of 0.80 and validation loss of 0.45.

Publication

Chaudhari, M., Gulati, I., Hundia, N., Karra, P., & Raval, S. (2025). *MoE Lens - An Expert Is All You Need.* Accepted at ICLR 2025 Workshop on Sparsity in LLMs (SLLM). URL: `https://openreview.net/forum?id=GS4WXncwSF`

Academic Service

Reviewer
ICLR 2026

Reviewer September 2025
NeurIPS 2025, Workshop on Multi-Turn Interactions in Large Language Models

Reviewer September 2025
NeurIPS 2025, Mechanistic Interpretability Workshop

Projects

CodeWhisper | *Python, TypeScript, Node.js, Deepgram API, Anthropic API, OpenAI API* October 2024

- Developed an VS Code extension enabling hands-free coding through natural speech commands
- Implemented real-time speech-to-code functionality by integrating Deepgram's speech recognition with language models using Groq's and Anthropic's API
- Designed an accessibility-focused system analyzing entire codebases for context-aware code editing and generation
- Built scalable architecture supporting multi-language coding assistance and real-time visual feedback
- Created at CalHacks, focusing on making coding more accessible for developers with visual or motor impairments

IdeaStruct | *Python, Flask, OpenAI API, Neo4j, Cytoscape.js* September 2024

- Developed an AI-powered Flask web app for generating and querying dynamic knowledge graphs
- Integrated OpenAI's GPT for NLP and implemented a flexible backend with Neo4j and in-memory databases
- Created an interactive frontend using Cytoscape.js for complex data visualization
- Engineered features like conditional data addition, URL scraping, and custom integration management
- Collaborated in a team of four to complete the project within HackMIT time constraints

Others: Textify: CNNs from Scratch, Glaucoma Prediction Model, etc.

Digital version available at www.idhant.xyz/resume