SURAJ ANAND

surajk610@gmail.com | (310) 987-0123 | surajk610.github.io

EDUCATION

Brown University

Providence, RI | May 2024

M.S. Computer Science, **B.S.** Computer Science–Applied Mathematics, GPA: 4.0

Thesis: *How to Promote Structural In-Context Learning with Forgetting*

Relevant Coursework: Grad Deep Learning (Python) · Grad Parallel Computing (Triton/CUDA/C++) · Grad Prescriptive Analytics (CPLEX/Go) · Grad Decision Making · ML Algos · Numerical Optimization (Matlab) · Statistical Applications · Information Theory · Compilers (OCaml) · Networks · Software Engineering

INDUSTRY EXPERIENCE

Point72/Cubist

New York, NY

Quantitative Researcher (NLP) — June - August 2023, August 2024 - Present

- Researched & productionized high-alpha NLP signals (IR > 1.5) in US equities using various large-scale data sources (TBs), scaled to $350M gross market value (GMV); continuous distribution shift testing.
- Constructed & combined mid-frequency signals to lower turnover and neutralize risk exposures.
- Owned LLM-based pipeline for structured knowledge extraction from internal notes & emails.
- **Embedding Research**: Trained/evaluated E5 & Qwen embedding models on synthetic QA data (4.3B tokens) with temporal consistency objectives to improve semantic alignment over evolving fiscal events; Improved NDCG@10 by 15 points on validation (*Torch, MTEB, Quantization, Tensorboard*).

Kaiser Permanente Medical Informatics

San Diego, CA

Machine Learning (NLP) Intern — May 2022 - August 2022

- Trained SBERT models on patient-reported pre-hospital visit reasons, improving downstream predictive performance by 10% (AUPRC) and enhancing robustness to edge cases *(SentenceTransformers)*.
- **Distributed Training**: Fine-tuned GPT-J (6B params) with ZeRO 3 parallelism to extract kidney stone features from radiology notes, addressing annotation scarcity through self-learning and active learning *(DeepSpeed, Torch)*.

RESEARCH EXPERIENCE

Brown University LUNAR Lab

Providence, RI

Research Assistant, Interpretability — 2022 - 2024

- **Pioneered structural in-context learning**: Developed a novel method for controlling when language models use in-context vs. in-weights learning called "temporary forgetting", tested on various tasks with toy models and GPT-2; published conference paper at ICLR.
- **Learning dynamics of syntax in MLMs**: Probed layers of the MultiBERTs for syntactic information (POS, NER, Phrase Start/End, etc) across 28 training checkpoints; found linearly-extractable representations pushed to earlier layers over training depending on data distribution.
- **Concept intervention & ablation**: Conducted systematic comparison of concept removal methods (INLP, RLACE, WTMT) on a pretrained CLIP-ViT with controlled visual datasets; demonstrated effectiveness of concept ablation and concept alteration algorithms using probe accuracy and downstream task performance.

PUBLICATIONS & PROJECTS

*Additional projects here

Controlling Use of In-Context vs. In-Weights Strategies with Weight Forgetting — **ICLR 2025**
Suraj Anand, *Michael Lepori, Jack Merullo, Ellie Pavlick*
Introduced structural in-context learning and weight forgetting techniques to control learning in transformers

RLAIF to Avoid Entities when Explicitly Instructed (Pink Elephants) — Preprint
*Louis Castricato, Nathan Lile, **Suraj Anand**, Hailey Schoelkopf, Siddharth Verma, Stella Biderman*
Applied RLAIF to solve instruction-following failures in language models in colab with EleutherAI

Jailbreaking PPO-ed Language Models with Mechanistic Interpretability — Preprint
Used mechanistic interpretability to reveal and exploit vulnerabilities in aligned GPT-2

Parallel simulated annealing for optimization (OpenMP/CUDA Kernel) — Presentation
Engineered and profiled high-performance parallel optimization algorithms

TEACHING & TECHNICAL SKILLS

Teaching Assistant, Computational Linguistics (CS1460) — *Providence RI* | Fall 2022 & Fall 2023

- Instructed 100+ students on Transformers, Word2Vec, Dependency Parsing, etc; developed final projects for Question Answering with BERT (2022) & Tracing Gender Bias in GPT2 with Mechanistic Interpretability (2023)

Deep Learning & NLP: *Torch, Transformers, Tensorboard, DeepSpeed, Ray, PEFT/LoRA, RLHF, FSDP, DDP, CUDA/Triton*
Data & Systems: *Python, SQL, C++, Go, OCaml, Spark, Docker, LanceDB, MTEB, Statistical Analysis, HPC*
Interpretability: *TransformerLens, NNSight, Logit Lens, Causal Interventions, Activation Patching, SAE Lens*

Achievements: AIME Qualifier (x2), Intel International Science Fair Finalist, National Merit Finalist