Chinmaya Andukuri

andukuri@stanford.edu | LinkedIn | Personal Website | GitHub | Google Scholar

Education

Stanford University March 2023 – . . .
M.S. in Computer Science (on leave, 28/45 units complete) *Stanford, CA*

Stanford University September 2019 – June 2024
B.S. in Mathematical and Computational Science *Stanford, CA*
Relevant Coursework: Deep Learning, NLP w/ Deep Learning, Deep Generative Models, Bayesian Statistics

Technical Skills

Programming Languages: Python, SQL, basic Rust
Technologies and Frameworks: PyTorch, vLLM, transformers, hydra, Weights + Biases, Git, Kubeflow, Triton

Experience

Software Engineer August 2024 – . . .
Capital One (Applied Research + LLM Pretraining Group) *San Francisco, CA*

- Built simulation + evaluation factory for multi-agent systems with LLM-as-a-judge and algorithmic metrics
- Enabled team to define tasks, generate synthetic seed data, and simulate conversations with LLMs for evaluation
- Improved LLM product's entity recognition + understanding accuracy by 20% using only synthetic data
- Reduced model size by 88% in production system with synthetic data while maintaining benchmark performance
- Constructed internal LLM leaderboard to automatically submit Kubernetes PyTorch jobs for model benchmarking

Student Researcher December 2023 – June 2024
Stanford Artificial Intelligence Laboratory (Computation & Cognition Lab) *Stanford, CA*

- First-authored COLM 2024 conference publication STaR-GATE on clarification and grounding
- Studied elicitation of preferences by language models through bootstrapping, simulation and self-improvement
- Built reusable repositories to study code problem-solving and reasoning abilities of language models

Software Engineer Intern June 2023 – August 2023
Capital One (Enterprise Data + Machine Learning) *McLean, VA*

- Constructed large language model (LLM) pipeline to provide search capability across enterprise
- Enabled $6 million in estimated savings for HR by embedding >7000 internal documents for semantic search
- Achieved 84% BERTScore F1 similarity between predicted and reference answers on open question-answering tasks

Software Engineer Intern June 2022 – September 2022
Dataherald, YC W21 *Los Angeles, CA*

- Implemented version control system module using Python/Git for MongoDB database with 400+ documents
- Created 20+ self-sufficient data pipelines using Databricks/PostgreSQL to create data visualizations for web app

Research Publications + Projects

Research Interests: Synthetic evaluation + training data, LLM bootstrapping / self-improvement, simulation

STaR-GATE: Teaching LLMs to Ask Questions (COLM 2024) | *vLLM, hydra* March 2024 – October 2024

- Developed a self-bootstrapping method to teach LLMs to ask better clarifying questions in multiturn conversations
- Trained `mistral-7b` and `llama3-8b` models to elicit information using expert response log-probs as reward signal
- Achieved 73% preferred response win rate over baseline instruction-tuned model

printllama | *PyTorch, vLLM* December 2023 – January 2024

- Built benchmark of 632 buggy code solutions to humaneval by sampling errors from abstract syntax trees (ASTs)
- Designed conditions to test whether print statements can improve LLM's bug-patching abilities
- Improved patch accuracy by 15% in `mixtral-8x7b` by allowing a pre-selection stage for LLM-preferred prints

FasterDecoding/REST (open source contribution) | *Rust, PyTorch* October 2024 – November 2024

- Localized byte-level implementation bugs in Rust-based open source retrieval tool for faster inference
- Enabled Llama 3 compatibility with REST framework by fixing integer bit-widths and PyTorch KV-caching issues

manipulativeLMs: Social Reasoning in LMs | *transformers, LoRA* November 2023 – December 2023

- Finetuned Stanford Alpaca-style language model to improve social reasoning ability
- Constructed 1000-example synthetic benchmark to test manipulative behavior in base- and finetuned- models