Pranav Saran

(510)709-7661 | pranav.saran@case.edu | LinkedIn | Github | Google Scholar

EDUCATION

Case Western Reserve University **Cleveland, OH**
- Major: Computer Science and Engineering 4.0 GPA
- Deans High Honors List Fall '24, Spring '25

TECHNICAL EXPERIENCE

Machine Learning Researcher – Palo Alto, CA January 2025 - June 2025
Algoverse (Python, PyTorch, Overleaf, LaTeX)
- Utilized a hybrid mBERT+BiLSTM model for figurative language detection, trained on low-resource Konkani
- Achieved an accuracy of 83% for idiom classification and 78% for metaphor classification
- Preserved 100% of original accuracy on idiom classification while pruning attention heads
- Achieved 88% accuracy retention on metaphor classification tasks through strategic model parameter reduction
- Performed ablation testing across multiple transformer architectures (mBERT, IndicBERT, XLM-R) to evaluate robustness and comparative performance

Software and Electronics Developer – Cleveland, OH September 2024 - Present
Case Western Reserve Global Health Design Collaborative (Arduino, C++, Python, Kivy)
- Engineered a Soft Access Point utilizing the ESP32 C3, facilitating seamless data transmission from the MAX30102 sensor to a mobile application developed in Python, enhancing real-time monitoring capabilities by 40%.
- Contributed to award-winning research showcased at CWRU Intersections, earning Second Place in Undergraduate Engineering, with findings from the project published and presented to a multidisciplinary academic audience; named as 2024-2025 team MVP
- Developed and integrated a mobile app using the KIVY library, enabling users to visualize heart rate and oxygen saturation data(SPO2) collected from over 500 sessions, thereby improving user engagement metrics by 60%.
- Collaborated with a cross-functional team to design and implement a detrending algorithm, improving data accuracy by 40% and enhancing the overall reliability of readings across a plethora of data points

PUBLICATIONS & RESEARCH ACTIVITIES

Pruning for Performance: Efficient Idiom and Metaphor classification in Low-Resource Konkani Using mBERT
Timothy Do, **Pranav Saran**, Harshita Poojary, Pranav Prabhu, Ivory Yang, Sean O' Brien, Vasu Sharma, Kevin Zhu
In *The 2025 Conference on Language Modeling, Workshop in Multilingual Data Quality Signals (COLM 2025 Workshop Paper, Still under review AACL 2025)*

- Served as Reviewer for: COLM WMDQS 2025

PROJECT EXPERIENCE

Transformer Implementation | Baby-GPT *Python, PyTorch, Transformers*
- Implemented the transformer based architecture from the paper *Attention Is All You Need*
- Developed a character tokenizer to encode text into tokens
- Implemented Self-Attention, Multi-Headed Attention and Normalization to train the Bigram Model on Shakespeare's works
- Replicated the performance of GPT-2 with successful implementation

Multi-Task Clinical NLP Pipeline | *BERT, HuggingFace, MIMIC-III*
- Preprocessed 40K+ clinical notes from MIMIC-III to create multi-label task formats
- Fine-tuned BioBERT with custom multitask heads, achieving 91% ICD F1 and 89% NER accuracy
- Used ROUGE metrics to benchmark summarization quality; improved performance by 18% over single-task baselines
- Applied attention head pruning to reduce model size by 30% while retaining 95% performance

Auto README Generator Agent *Python, SmolAgents, OpenAI GPT-4o, Gradio*
- Developed an autonomous agent using smolagents to automate technical documentation by analyzing codebases and generating READMEs using LLM-driven reasoning
- Engineered a modular toolchain for code parsing, dependency inference, and capability extraction, enabling structured analysis of arbitrary GitHub repositories
- Integrated a Gradio-based interface for seamless human-agent interaction, showcasing system design, prompt engineering, and practical application of multi-tool LLM workflows

Autonomous Agent for Reasoning Benchmarks *Python, smolagents, OpenAI API, LLM Tooling*
- Developed and deployed a multimodal reasoning agent using the smolagent framework, achieving automated task-solving across search, code, and file understanding using tools like Wikipedia lookup and YouTube analysis.
- Integrated GPT-4o with custom tools and evaluated agent performance on the GAIA benchmark via Hugging Face Spaces, using REST APIs to batch-run and submit results.
- Engineered tool schemas, error handling, and agent orchestration logic to improve generalization and robustness across diverse natural language and data-driven tasks.

TECHNICAL SKILLS

- Programming Languages: Python, C++, Java, Javascript
- Frameworks: PyTorch, Flask, Node.js, Express.js, MongoDB, Firebase, Kivy, JFrame, Chrome API, NEAT, Linux
- Relevant Coursework: Data structures, Computer Security, Discrete Mathematics, Logic Design and Computer Organization, Linux and OS
- Certifications: The LLM Course (Hugging Face), Fundamentals of MCP (Hugging Face), AI Agents Course (Hugging Face)