Pradham Mummaleti

☎ (352)-278-1451 ✉ pradhammummaleti@ufl.edu linkedin.com/in/pradhammummaleti github.com/cxrlton

Education

University of Florida **Aug 2024 – Dec 2025**
Master of Science in Computer Science GPA: 3.76/4.0 *Gainesville, Florida*

Mahindra University, École Centrale School of Engineering **Aug 2020 – May 2024**
Bachelor of Technology in Artificial Intelligence *Hyderabad, Telangana*

University of Florida **Jan 2024 – May 2024**
CISE Exchange Student *Gainesville, Florida*

Publications

- W. Coggins, J. McKenzie, S. Youm, **P. Mummaleti**, J. Gilbert, E. Ragan, and B. J. Dorr. "*That Ain't Right: Assessing LLM Performance on QA in African American and West African English Dialects.*" In *Proceedings of the 9th Widening NLP Workshop*, ACL 2025, Suzhou, China.
- C. Srikanth, S. R. Yerabelly, **P. Mummaleti**, and A. Jain. "*Automation of Human Body Measurements from 2D Images.*" In *Proceedings of the 5th International Conference on Frontiers in Computing and Systems (COMSYS-2024)*, December 2024.

Research Experience

GPT2 - PyTorch Reimplementation **Dec 2024 (Link)**

- Reimplemented GPT-2 (∼124M) in PyTorch, trained on FineWeb-Edu (10B tokens) and TinyShakespeare datasets using DDP for distributed training across 4x NVIDIA L40S GPUs.
- Integrated optimizations including Flash Attention (6x faster), `torch.compile` for JIT-compiled kernels, and mixed precision for enhanced training throughput.
- Achieved a final validation loss of 0.0372 on the TinyShakespeare dataset and scaled compatibility with large datasets such as FineWeb-Edu (10B tokens).

Implementation of Memorizing Transformers with kNN-Augmented Attention **Jan 2024 (Link)**

- Implemented the Memorizing Transformers architecture with kNN-based memory retrieval and external memory management, handling sequences up to 5120 tokens using PyTorch and FAISS.
- Trained the model on T4 GPUs with a pipeline incorporating Transformer-XL recurrence and T5 positional encodings, optimized for datasets like arXiv Math with memory sizes up to 81,920 tokens.

DeepResearch – Multi-LLM Semantic Orchestration Framework **Oct 2025 – Dec 2025**

- Explored how collaborative orchestration among small language models can replicate large-model reasoning, deploying on UF's HiPerGator GPU cluster via vLLM for scalable inference.
- Curated cross-task benchmarks to evaluate faithfulness, depth, latency, and cost across SLM–LLM setups.
- Designed a semantic routing and critique pipeline combining local (LLaMA-3.1-8B, Mistral-7B, Gemma-9B) and frontier models (GPT-4o, Claude 3, Gemini 1.5 Pro), improving reasoning accuracy 15% while cutting cost 70%.

Technical Skills

Languages: Python, C/C++, Julia, Pony, R, SQL
Tools & Frameworks: NumPy, Pandas, Scikit-learn, PyTorch, Jupyter, Tensorflow, AWS Web Services, Hugging Face

Certifications

Senior Certificate in Computer Science & Engineering | *University of Florida* **Jan 2024 - May 2024**

- Completed the final undergraduate semester through an exchange program at the University of Florida, earning 9 graduate-level credits.

Generative AI with Large Language Models | *Coursera, Amazon Web Services, DeepLearning.AI* **Oct 2023**

Natural Language Processing with Classification and Vector Spaces | *Coursera, DeepLearning.AI* **June 2024**

Awards

- **Mahindra University Merit Scholarship** — awarded for academic excellence in AY 2020–2021, 2022–2023, and 2023–2024.
- **University of Florida CISE Graduate Scholarship** — awarded for outstanding academic performance.