D Ravi Theja

LinkedIn | Github | ravi03071991@gmail.com | +1-6503803923
Medium Blog

Experience

MistralAI Palo Alto, CA
Applied AI *Feb 2025 - Present*

- Developed enterprise AI systems using Mistral LLMs, OCR models, and embedding models to automate document and knowledge workflows.
- Designed evaluation datasets derived from real enterprise customer workflows, identifying OCR model failure cases and collaborating with modeling teams to improve model robustness.
- Reduced invoice reconciliation time by 99% (2 hours → 1 minute) by architecting an LLM-driven document understanding pipeline with structured extraction and code-mapping workflows.
- Built multi-agent systems for automated PRD generation from meeting transcripts, earnings call analysis, financial reporting, and industrial knowledge retrieval. Blog | Cookbooks

LlamaIndex Remote
AI Engineer And Developer Advocate *Oct 2023 - Jan 2025*

- Developed evaluation modules for RAG systems and introduced the use of LLM-as-judge for evaluating retrieval quality, while integrating multiple data loaders for efficient ingestion.
- Implemented latest RAG research papers - GraphRAG, CorrectiveRAG, AdaptiveRAG, and Mixture Of Agents as LlamaPacks.
- Developed document-driven agentic applications to enrich invoice processing through LlamaIndex workflows.
- Developed an Orielly Media course on Building RAG Applications with LlamaIndex. Course
- Integrated latest LLMs, embedding models, rerankers, and fine-tuning APIs from OpenAI, Anthropic, Mistral, Cohere, and provided essential support for issues and PRs within the OSS framework.
- Conducted analysis and benchmarking for RAG systems, significantly improving LlamaCloud's retrieval performance on complex queries through sub-query planning and metadata filtering.
- Advised RAG solutions for clients such as ByteDance, EY, NetApp India, Albus, Atomic Works, and Videoverse

AI Researcher Bangalore, India
Part-time - Independent OSS Work *Jan 2024 - Aug 2024*

- Integrated MMMU benchmark testing capabilities into SGLang LLM inference library, enabling standardized evaluation of Vision-Language Model performance. PR
- Developed Navarasa 2.0, a Gemma-based multilingual LLM supporting 15 Indian languages, featured in Google's Gemmaverse and showcased at Google I/O 2024. Gemmaverse, Blogpost, Models, Code.
- Developed Navarasa 2.0, a Gemma-7B/2B finetuned model for 15 Indian languages, featured in GoogleIO keynote 2024. Blogpost, Models, Code.
    1. Managed end-to-end aspects of model creation including data preparation, modeling, and evaluation.
    2. Positioned Navarasa 2.0 among the top-6 models for various Indian languages as evaluated by Microsoft Research. Undertook this project as part of collaborative work under TeluguLLMLabs.
- Developed the BRAG series of Small Language Models (SLMs) optimized for RAG, surpassing performance benchmarks of major models like Cohere's Command R+, Qwen2, Llama3.1, and Llama3 Instruct, and closely matching GPT-4-Turbo and Nvidia's ChatQA-1.5-8B on Nvidia's ChatRAG-Bench. Blogpost, Models.
    1. Enhanced model performance through strategic data mix and minimal dataset exploitation using LoRA and QLoRA technologies.
    2. Achieved cost-effective model training, with each model developed under $25. Undertook this project as part of collaborative work under maximalists.ai.
- Automatic Knowledge Transfer (KT) Generation for Code Bases: A project to streamline knowledge transfer in IT by summarizing and explaining code, then transforming these into engaging videos using LlamaIndex, D-ID's text-to-speech, and video generation. These integrations make complex codebases more accessible and simplify onboarding processes. Blogpost, Code

- Global NeurIPS Paper Implementation Challenge: Implemented the NeurIPS-2017 paper "Selective Classification For Deep Neural Networks," which introduces a selective classifier for CNNs to determine instance reliability before prediction, requiring human input when uncertainty is high. Code

Glance - Inmobi Bangalore, India
Senior Machine Learning Engineer *March 2021 - Oct 2023*

- Developed and deployed the Glance TV Screen Saver product to automate wallpaper creation for the latest news articles on TV, including generating concise headlines and sub-headlines from text, utilizing CLIP Embeddings and Sentence Transformers for image searches, and stitching text with images for display.
- Achieved an 85% reduction in content creation time (20 min → 3 min) by designing and implementing an AI-assisted pipeline for Glance TV Screen Saver product.
- Developed and deployed GPT-3 based systems on Glance TV, introducing a paraphrasing comment generation system and an automated poll generation framework. These systems boosted user interaction, increasing watch time by 32.07% and reducing the workload for content and editorial teams by 30%.
- Developed and deployed deep learning based recommendation models, including a DropoutNet and an auto-encoder with ALS-based recommendation model, which significantly improved interaction rates, impressions, and user engagement duration for cold and sparse user segments.

TCS Innovation Labs Pune, India
Research Engineer *Aug 2019 - March 2021*

- Developed attention mechanism architecture for detection of humor in edited news headlines using BiLSTM, knowledge graph and other extracted features. Paper published at COLING - 2020 conference in SEMEVAL workshop.

Quadratic Insights Pvt. Ltd. Hyderabad, India
Data Scientist *Mar 2016 - Jun 2017*

- Developed a hierarchical algorithm using text mining techniques and Naive Bayes algorithm to automatically redirect the customer complaint emails of a leading bank to the respective departments at three hierarchical levels.

Hindustan Petroleum Pvt. Ltd. Vijayawada, India
Operations Officer *Jul 2013 - Jun 2014*

- Validated machine learning models and developed new features to forecast sales of different oil products that helped in running the distribution plan effectively.

Education

International Institute Of Information Technology, Bangalore (IIIT-B) Bangalore, India
Master of Science in Computer Science; GPA: 3.77/4.0 *July 2017 - 2019*

National Institute Of Technology, Warangal (NIT-W) Warangal, India
Bachelor of Technology in Electrical And Electronics Engineering; GPA: 8.14/10.0 *Jul 2009 - Apr 2013*

Achievements

- Presented Navarasa-2.0, a Gemma finetuned model for 15 Indian languages, at the GoogleIO keynote session-2024.
- Winner of the Google Cloud, Searce, and LifeSight hackathon for our Automatic Knowledge Transfer (KT) system for code bases.
- Winner of the Google Award at the LightSpeed hackathon for developing parah.ai, an AI-driven tutoring system.
- Published research paper on summarizing short medical conversations at ACL-2023 MEDIQA-Chat.
- Presented Automatic Knowledge Transfer (KT) Video Generation Of Code Bases using LlamaIndex at PyCon India-2023.
- Named among the top-5 GenAI Professionals in India by Analytics Vidhya at the Data Hack Summit.
- Published research paper on Humor Recognition at COLING-2020's SemEval-2020 workshop.
- Included in the Dean's Merit List during the Master's program at IIIT-B.
- Provided AI advisory services by collaborating with the founders of the startup Composio to enhance the developer experience.
- Winner of the Global NIPS Paper Implementation Challenge, 2017.

Skills

- Python, LLMs, NLP, LlamaIndex, SQL, Deep Learning, Recommender Systems.