YASH BHASKAR

Student at IIIT Hyderabad

Hyderabad, India | yash9439@gmail.com | yash9439.github.io | Linkedin | Github | Medium

ABOUT

I am a student at IIIT Hyderabad specializing in AI and NLP. I have worked with many organizations on projects involving data ingestion for large language models, code query systems, and AI research. I love working on new and challenging tasks. In my free time, I enjoy writing articles and listening to music. I am passionate about advancing AI technologies and contributing to the field through innovative research and practical solutions. My aim is to make a significant impact in the AI community.

EDUCATION

B.Tech in Computer Science
MS in Computational Linguistics

IIIT Hyderabad

July 2021 – Present

- AI/NLP specialization

EXPERIENCE

Binoloop — Remote - Canada

Research-Intern — Oct 2024 – Nov 2024

Legal Document Evaluation Pipeline

- Improved the Legal Document evaluation pipeline using Binoloop's proprietary **Tally Framework**, focusing on enhancing relevant chunk retrieval for more accurate document analysis.

Epicurist Lab Inc. - ClearEdge.ai — Remote - US, California

Freelance - Contract — March 2024 – Present

Clearedge - Data ingestion for LLMs

- Developed **Clearedge**, an open-source Python library for LLM data ingestion, implementing text and metadata extraction from complex scanned pdf and other file formats.
- Analyzed document segmentation and OCR solutions like Tesseract, PaddleOCR, DocTR, Microsoft TableBank and YoloV9. Implemented Clearedge's core pipeline for text, image, and table data extraction.

Code Indexer

- Created a code query system for large repositories using tree-sitter, enabling semantic analysis of functions across multiple files.
- Implemented a code snippet ranking system using PageRank and leveraged GPT-4o for response generation. I designing the core logic and deployed its backend.

SuperTeams.ai — Remote - Delhi, India

AI Researcher and Writer — Feb 2024 – Present

- Partnered with Qdrant and Astria.ai for technical content creation, producing in-depth articles on their platforms' applications in enterprise cloud-based RAG systems and virtual staging using custom diffusion models.
- Made articles on RAG Pipelines, Vector Databases, Distributed RAG System using RAY, ReAct Agents in RAG, Documentation RAG System with Internet Access.

ISB — Remote - Delhi, India

Research Intern — May 2024 – July 2024

- Conducted research on footfall data for various businesses in North America, analyzing patterns based on political timelines and festival calendars.
- Developed and implemented data scraping pipelines to collect relevant datasets.
- Utilized Google Cloud Geocaching API to obtain latitude and longitude data, and Bing Distance Matrix API to calculate distances for spatial analysis.

SUBTL.AI — On-site - Hyderabad, India

Research Intern — May 2023 - July 2023

- Built a validation layer, flagging a portion of output for enhanced correctness checks.
- Redesigned the PDF Parser for accurate text detection, resulting in a reduction in errors.

RCTS - Center — On-site - Hyderabad, India

Full Stack Web Developer — Dec 2022 – April 2023

- Built a user-friendly crowd-sourcing platform integrated with GitLab, allowing project assignments and contributions. Simplified collaboration among developers.
- Developed an interface for website admins to use GitLab features through clicks, eliminating the need for coding. Streamlined project management and code deployment processes.

PUBLICATIONS

Assessing Translation Capabilities of Large Language Models involving English and Indian Languages - EAMT24 Conference

- Explored and optimized multilingual machine translation capabilities in large language models, focusing on English and 22 Indian languages, using methods such as in-context learning, LoRA, and full fine-tuning, based on the LLaMA model.
- Improving multilingual translation in large language models for English and 22 Indian languages using ICL, LoRA, and full SFT.

Yes-MT's Submission to the Low-Resource Indic Language Translation Shared Task in WMT 2024

- Explored translation using methods like LoRA fine-tuning of large language models, and fine-tuning of models such as IndicBART and mT5.

Fine-tuning Language Models for AI vs Human Generated Text detection - NAACL24 Conference

- Developed a RoBERTa-based system for detecting machine-generated text.

Disfluency Detection using Neural Models - ICON23 Conference

- Fine-tuned a model for disfluency detection in spoken text across six Indian languages.

KEY PROJECTS

Amazon ML Challenge 2024

- My team achieved an All India Rank (AIR) of 15 out of 75,000+ participants in the Amazon ML Challenge 2024 organized on Unstop.
- OCR Pipeline: Used OCR models like DocTR and PaddlePaddle for text extraction, followed by a prompt-tuned Llama 3.1 instruct model (7B) to retrieve entity values from the extracted text.
- Visual-Language Models: Leveraged Qwen-VL models (2B and 7B) to directly extract entity values from images, with additional rule-based text processing for enhanced accuracy.

RAG-with-Agents-llama3 — yash9439/RAG-with-Agents-llama3

- Developed an intelligent information retrieval system integrating LangChain's ReAct Agents, Qdrant vector database, and Llama3 language model. Created custom tools for agents including document retrieval, text summarization, date lookup, and age calculation to enable complex query processing and information extraction.

Code-Mixed-Machine-Translation — yash9439/Code-Mixed-Machine-Translation

- Developed models for English to Hinglish Translation. Compared performance among fine-tuned mT5, IndicBART, Mistral-7b, Indic-Trans-2 and Meta Seamless.

Prompt-Tuning-GPT2-Pytorch — yash9439/Prompt-Tuning-GPT2-Pytorch

- Explored prompt tuning techniques for fine-tuning GPT-2 in PyTorch. This project involved experimenting with various prompt tuning methods to fine-tune GPT-2 for specific tasks or domains, enhancing its performance and adaptability to different use cases.

Reddit-Clone — yash9439/Reddit-Clone

- Designed and implemented a comprehensive Full Stack Reddit Clone using the MERN (MongoDB, ExpressJS, ReactJS, NodeJS) Stack and MUI Library. Features include post creation, commenting, upvoting/downvoting, user following, and subgreddit moderation.

CPU-Scheduler — yash9439/CPU-Scheduler

- Developed a CPU scheduler simulator in C to visualize and analyze the performance of common scheduling algorithms (FCFS, SJF, Priority, Round Robin). Implemented data structures and algorithms to simulate process execution, calculate waiting times, and provide a user-friendly interface.

Chat-With-Multiple-Pdfs — yash9439/chat-with-multiple-pdf

- Developed a system for interactive conversations with PDF documents using LangChain, ChromaDB, and OpenAI's API. This project enabled users to engage in interactive conversations with PDF documents by extracting and processing text using LangChain, managing document metadata with ChromaDB, and leveraging OpenAI's API for natural language understanding and response generation.

RNN-LSTM-GRU-Transformers-Pytorch-Implementation — yash9439/NeuralNetArch-PyTorch

- Implemented Multi-Layer RNN, Multi-Layer LSTM, Multi-Layer GRU and Transformers from scratch using torch.nn only.

ACHIEVEMENTS

- **Content Creation:** Authored 80+ technical articles on AI, ML, and Language Models, amassing over 200,000 views on Medium.
- **Technical Writing Partnerships:** Partnered with Qdrant to create technical articles on building enterprise cloud-based RAG systems.
- **Dean's List Awardee and Research's List Awardee** for 2024 Spring Semester for excellence in academic performance
- **Academic Mentorship:** Project Mentor at IASNLP Summer School and Speaker at IASNLP Preschool, 2024.
- **Conference Presentations:** Presented research works at multiple prestigious conferences including ICON 2023, NAACL 2024, and EANMT 2024.
- **Ranks: JEE Advanced: Secured AIR 1892**, Class XII: 95.6%, Class X: 97.6%