Nirmal Thomas

✉ nirmal.thomas013@gmail.com • 🌐 nrmlthms.github.io
in nirmal-thomas • nrmlthms

Education

Noida International University **Noida, UP, India**
BSc Data Science (CGPA: 8.41) *Sep 2021 – June 2024*

Work Experience

Pratham International **Remote – Mumbai, India**
AI Engineer *Sep 2024 – Present*

- Designed and deployed multiple RAG-based chatbots including:
  - **BaalSakhi** – Deployed on WhatsApp and trained on documents from government departments, UNICEF, WHO, Pratham, pediatricians, and childcare experts; supports mothers in low-resource communities by answering early childhood care queries in regional languages through both text and audio interactions.
  - **ASER Bot** – Developed for India's largest citizen-led education survey providing reliable data on learning outcomes to enable real-time access to ASER data and insights for policymakers and educators using a Hybrid-RAG approach.
  - **Pratham Website Chatbot** – Built a website assistant to help users navigate organizational information efficiently.
- Built speech-to-speech conversational bots using cascaded pipelines, deployed for skilling programs in India and Kenya.
- Built a Voice-based Surveyor for facilitators implementing the Teaching at the Right Level (TaRL) methodology, an AI assistant that identifies evidence-based nudges and personalized pedagogical recommendations from prior facilitator interactions.
- Designed and implemented an in-house translation and transcription pipeline, reducing production costs by over 80% for multilingual video and subtitle generation.

PAiGPT.ai **Noida, UP, India**
AI Engineer *Sep 2023 – August 2024*

- Built an Answer Engine leveraging LLM agents, advanced prompting, and real-time crawled data for precise, context-aware responses.
- Fine-tuned a pre-trained LLM using Unsloth to accurately generate source citations and improve factual reliability.
- Designed and implemented complete RAG architecture to improve domain adaptation of the proprietary in-house LLM.
- Utilized a vector database to power a scalable recommendation system integrated with the chatbot ecosystem.
- Introduced Hindi text generation using a custom translation pipeline built with an NMT model and CTranslate2, enabling multilingual support.
- Created a data ingestion pipeline using Change Data Capture (CDC) to integrate unstructured data from PDFs and web sources.
- Developed production-grade backend APIs with authentication, rate limiting, and other reliability features.
- Implemented comprehensive security and reliability measures, including prompt injection defense, output guardrails, prompt caching, data encryption, API authentication, and retrieval validation.

Jishu Excellence **Remote - India**
Data Science Intern (Part Time) *July 2022 - July 2023*

- Assisted in building an Audio Intelligence System that converts call recordings into diarized text with emotion and sentiment analysis to identify prospective customers based on defined metrics.
- Improved overall system reliability by **12%** through evaluation-driven optimization and metric-based feedback loops.
- Enhanced sentiment and emotion detection accuracy by **4%** using integration of state-of-the-art NLP models.
- Developed a Chatbot interface to interact with transcriptions, extract insights, and perform one-click summarization using transformer-based models.

Research Experience

Summer of Open AI Research **Hosted by EleutherAI**
Prompt Optimization for Hallucination Reduction *August 2025*

Expedition Aya 2025 **Hosted by Cohere Labs**
Multilingual Speculative Decoding *April 2025 - May 2025*

Publications, Theses and Pre-prints

MRL Workshop Shared Task, EMNLP 2025: **Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures** [Preprint]
Contributed to the construction of the dataset in Hindi.

Bachelors Dissertation: **Decoder Trimming - A Feasible Alternative to Fine-Tuning for Text Classification with Large Language Models**
Advised by Dr. Aashima Bangia, Investigated the impact of decoder trimmed LLMs for classification of text.

Awards and Conferences

May 2025: **Expedition Aya: Most Promising Award** – Explored multilingual speculative decoding to make LLMs more efficient.

February 2025: **Best Student Award** – Awarded for overall outstanding performance during undergraduate studies.

Skills and Interests

Tools and frameworks: PyTorch, TGI, Transformers, HuggingFace, vLLM, FastAPI, Flask, LangGraph, Ctranslate2

Interests: Efficient training/optimization methods and inference, transformers, large language models, Multilingual, Compression, Information Retrieval