Prathamesh Mahale

AI Engineer | LLM Applications | Semantic Search | NLP Systems

8262032848 | prathameshmahale734@gmail.com | github.com/beweme

Profile

AI Engineer with experience building and deploying LLM-powered products, semantic search systems, recommendation infrastructure, and NLP pipelines at production scale. Proven track record in shipping retrieval, ranking, summarization, and content understanding systems across catalogs with millions of items, while improving latency, cost efficiency, and workflow automation. Strong foundation in applied machine learning, deep learning,LangGraph, FastAPI, PyTorch, Transformers, and cloud-based ML orchestration across AWS and Vertex AI.

Experience

AI Engineer, VP of Tech's Office Sept 2025 – Present
Pratilipi *Bengaluru, India*

- Built **Hubble**, an LLM-powered semantic search engine using **LangGraph** over **500K+ stories** to match high-potential IP with adaptation briefs from OTT platforms including **Netflix, Amazon Prime Video, and SonyLIV**, reducing content discovery time from **4 hours to 30 minutes** and driving a **2× increase** in IP exploration.
- Developed an **LLM-based manuscript triage pipeline** processing **600K–800K words per week**, generating structured summaries and editorial signals for **1,500+ submissions**, reducing review latency by **50%** and saving **12.5 hours/week** of manual effort.
- Built a **platform-wide embedding infrastructure** for a **5M+ story catalog**, generating reusable semantic vectors for search, recommendation ranking, and diversity monitoring, while reducing embedding cost by **60%** using **Vertex AI Batch** pipelines.
- Work directly with **VP of Technology** to **consult, architect, and deploy production-grade LLM and agentic systems** across the organization, owning end-to-end delivery from ambiguous problem statements to reliable systems in production.

Data Scientist Intern Mar 2025 – Jun 2025
Pratilipi *Bengaluru, India*

- Implemented **SVP-CF**, a data sampling strategy for recommendation systems, reducing training data volume by **60%** while preserving **A/B test performance**, improving model training efficiency at scale.
- Refactored the internal **Facebook ad generation pipeline**, improving modularity, logging, and observability, and reducing **OpenAI API usage by 60–80%** through caching and reuse of prior generations.
- Migrated the pipeline to **AWS Step Functions** using **Metaflow**, enabling parallel execution and increasing workflow throughput by **6×**.
- Contributed to an **AI story-writing agent** using **LangGraph** that generated narrative drafts **35× faster** and **10× more cost-effectively** than manual writing workflows.

Artificial Intelligence Research Intern Nov 2024 – Mar 2025
University of Arkansas, Little Rock *Little Rock, AR*

- Improved biomedical NER performance by **4% balanced accuracy** through **weighted cross-entropy loss**, optimizing minority-class learning under severe label imbalance.
- Designed a **knowledge distillation pipeline** from **BioBERT** to a lightweight student model using **soft-logit supervision** and word-level logit alignment, improving **F1 score and balanced accuracy by 2–5%**.
- Integrated **focal loss** into the NLP training pipeline, achieving **state-of-the-art balanced accuracy** for the task.
- Built **2D and 3D embedding visualization workflows** for BERT-family models, accelerating error analysis and improving interpretability of model behavior.

Artificial Intelligence Research Intern Sep 2023 – Jan 2024
HCL Tech *Pune, India*

- Built an **OCR + LLM-based document understanding pipeline** to extract dimensional rules and structured information from technical PDFs in the hardware manufacturing domain, improving model accuracy by **30%**.
- Fine-tuned **DistilBERT** on a custom rule-extraction dataset, achieving **98% test accuracy**.
- Created evaluation and visualization workflows over **10+ technical PDFs** to support extraction quality analysis and manual review.

Projects

PDF-to-Rule Converter for Manufacturing Insights | *Python, PyMuPDF, PyTorch, Transformers, Git*

- Developed and fine-tuned **LLM/NLP pipelines** to classify sentences containing dimensional rules for 3D models, improving classification accuracy by **48%**.
- Built a **LLaMA 3**-based prototype to parse extracted dimensional rules into structured JSON, achieving **78% label match** on an internal test set of **300 samples**.
- Reduced manual parsing workload by enabling an estimated **4× faster** review workflow for downstream teams.

LLM-powered RAG System for Airport Authority of India | *FastAPI, Python, Transformers, Hugging Face*

- Developed a **streaming text generation API** with **FastAPI** capable of serving concurrent users with average response latency of **300–500 ms**.
- Implemented **semantic chunking, clustering, and retrieval** over large text corpora using embeddings and KMeans, achieving **95% retrieval accuracy** for context-grounded queries.
- Optimized transformer inference with **4-bit quantization**, reducing memory footprint by **4×** and improving inference speed by **2×**.
- Built a customized **RAG pipeline** for question answering over Airport Authority of India content.

Legal Document Generation Engine | *LLaMA 3, ReactJS, TailwindCSS, FastAPI, LaTeX, REST APIs*

- Built a legal document generation platform with APIs for **6+ document types**, reducing drafting time by **70–80%**.
- Implemented backend document rendering with **LaTeX** and exposed generation workflows through **FastAPI** endpoints integrated with a React frontend.
- Developed an **LLM-powered chatbot** for legal information retrieval and assistance on the platform.

Technical Skills

Languages: Python, C++
AI/ML: PyTorch, Transformers, Langchain, Hugging Face, TensorFlow, NLP, LLMs, RAG, Semantic Search, Recommendation Systems, Knowledge Distillation
Backend & APIs: FastAPI, Django Rest Framework, REST APIs
Data & Analysis: NumPy, Pandas, Matplotlib
Cloud & MLOps: AWS (S3, Step Functions, Batch), Vertex AI, Azure Blob Storage, Docker, Git, Metaflow

Education

Vishwakarma Institute of Information Technology — Pune, India
Bachelor of Technology in Artificial Intelligence and Data Science — *Aug 2022 – May 2026*

Highlights & Achievements

Regional Industry Summit 2024

- Project selected for representation at the **Regional Industry Summit 2024**, organized by IFCCI, IGCC, and DSCI, where it was presented to **50+ industry leaders and startup founders** in Generative AI, Machine Learning, and data privacy.

Overall Winner – Vortexa Tech Hackathon 2024

- Won **first prize** among **250+ participants** by building a **Legal Document Generation Engine** and a **GenAI chatbot** in a **12-hour** hackathon.

Smart India Hackathon 2024 Grand Finalist

- Selected as a **Grand Finalist** among **380 teams** for building an AI-driven legal research prototype designed to accelerate legal information retrieval.