NIKHILESH GAWHALE

nikhileshgawhale711@gmail.com | +44 7767 941384 | LinkedIn| Github

SUMMARY

AI/ML Researcher with a Distinction MSc in Artificial Intelligence from QMUL, passionate about multimodal learning, affective computing, and explainable AI. Experienced in building and deploying NLP and LLM systems (e.g., LLaMA 2, GPT), with a strong foundation in deep learning, Bayesian methods, and ethical AI. Proven ability to bridge research and application through academic projects and industry deployments, with a focus on robustness and contextual awareness in real-world environments.

TECHNICAL SKILLS

- **Languages/Frameworks**: Python, TypeScript, Node.js, React, FastAPI, Flask, Scikit-learn
- **LLMs & NLP**: GPT, LLaMA 2, Generative AI, Prompt Engineering, RAG, NER, Summarization
- **Tools/Platforms**: Hugging Face, Docker, Kubernetes, Terraform, GCP, GitHub Actions (CI/CD)
- **Practices**: MLOps, REST APIs, Feature Engineering, Microservices, Data Visualization, Edge Deployment
- **Soft Skills**: Communication, Teamwork, Problem Solving, Adaptability, Critical Thinking

EDUCATION

Queen Mary University of London – *London, UK*
Master of Science in Artificial Intelligence (Distinction) *Sep 2023 – Sep 2024*
Key Modules: NLP, Deep Learning, Bayesian Neural Networks, ML Algorithms

Shivaji Science College – *Amravati, India*
Bachelor of Computer Applications (GPA: 7.8/10.0) *Aug 2019 – Aug 2022*
Key Subjects: Software Engineering, Data Structures, OOP, DBMS

EXPERIENCE

Generative AI Intern – *Midas Advisory, London, UK* *Aug 2024 – Nov 2024*

- Optimized LLM performance, reducing financial report analysis time by 25% through fine-tuning and prompt engineering.
- Developed an AI pipeline to process 10,000+ financial records daily, enhancing model adaptability with LLaMA 2.
- Enhanced RAG-based retrieval, increasing contextual relevance by 20% using Milvus for efficient knowledge augmentation.
- Implemented structured prompt templates, which improved financial insight accuracy by 18% and reduced hallucination rates.
- Increased system scalability by 40% and reduced model training costs by 30% through effective data preprocessing, hyperparameter optimization, and integrating Kubernetes-based microservices.

Software Developer-GLOBAL Technologies, Pune, India *Aug2022 – Aug2023*

- Built and deployed ML-powered debt automation system, improving response rates by 40%.Executed manual test cases for e-commerce platforms, ensuring timely, software delivery aligned with requirements.
- Developed test automation with predictive analytics, reducing defect rate by 50% and improving time-to-market..
- Integrated ML into mobile QA workflows, boosting test efficiency by 45%.

PROJECTS

Dissertation: *Exploring the Relationship Between Temperature and Creativity in Large Language Models.(Distinction)*
***Supervisor**: Prof. Massimo Poesio*

- Designed a novel experimental framework to evaluate creativity in LLMs using four task types: divergent thinking, convergent thinking, story continuation, and poetry generation Demonstrated 20% increase in output diversity, 15% improvement in relevance
- Benchmarked five LLMs (e.g. LLaMA3-70B, Mistral-7B, Phi3, Gemma-2B, TinyLlama) across multiple temperatures using a multi-metric creativity assessment paradigm (DAT, TTCT: fluency, flexibility, elaboration)
- Found that higher temperatures improve novelty and fluency but decrease coherence in open-ended tasks, offering implications for LLM deployment in creative or affective domains
- Insights contribute to developing adaptable and personalized AI systems that balance randomness, robustness, and task-specific performance

LLaMA 2 Chat Model Enhancement:

- Fine-tuned LLaMA 2 13B-chat model with domain-specific prompts
- Achieved 18% gain in generation accuracy and improved interactive performance
- Conducted extensive hyperparameter tuning using validation-based techniques
- Implemented conversational dataset curation strategies for model personalization
- Benchmarked outputs against GPT-3.5 using BLEU and ROUGE metrics

Entity Recognition System:

- Built NER model with dual GRUs in Keras; F1 score: 77.08%
- Added token-level audit tracing for compliance use cases.
- Integrated model into a Flask API for real-time deployment
- Deployed via FastAPI on GCP Cloud Run (serverless), simulating production traffic.
- Deployed on Kubernetes using CI/CD pipelines for version-controlled rollouts.

Dialogue Act Tagging Tool:

- Used Bi-LSTM for classifying conversational dialogue acts
- Boosted classification accuracy by 30% over baseline using domain-tuned inputs
- Augmented dataset using back-translation and synonym replacement
- Evaluated model using confusion matrix and class-wise recall metrics
- Created a simple GUI in Streamlit to visualize dialogue tag predictions

Sentiment Analysis on Tweets:

- Real-time tweet classifier using SVM, achieving 25% precision gain via feature engineering
- Performed text normalization and TF-IDF vectorization for preprocessing
- Built pipeline to stream tweets using Twitter API and classify in real-time
- Added dashboard for sentiment trends using Plotly and Dash
- Incorporated sarcasm detection features to improve accuracy in ambiguous tweets