Siddhant Gupta

+91 7836015759 | writetosiddhant@gmail.com | LinkedIn | Github | Google Scholar | Personal Website

Education

Indian Institute of Technology Roorkee — Roorkee, Uttrakhand
Bachelor of Technology — *Oct 2022 – July 2026*

Experience

Cohere For AI (C4AI) — June 2023 – Present
Research Lab and Open Science Community — *NLP Lead*

- Engaged in 50+ technical discussions and workshops on topics such as NLP, multi-agent systems, contextual learning, synthetic data generation, and mechanistic interpretability, contributing to the community's knowledge base.
- Led implementation efforts for research papers, collaborating with researchers globally to work on the latest methodologies mainly RAG, interpretibility, framework designing and Agentic systems.
- Worked on a 8-week long hackathon Expedition Aya where I developed speech synthesis method using ASR data.

Artificial Intelligence and Electronic Society (ArIES) — May 2023 – Present
Indian Institute of Technology, Roorkee — *ML Executive*

- Collaborated with cross-functional teams to participate in Inter-IIT competitions
- Spearheaded teams in AI hackathons, providing mentorship in CV and NLP research alignment, leading to the successful implementation of 10+ innovative projects
- Organized and conducted workshops and talks for 100+ participants, focusing on deep learning and image processing concepts such as edge detection, depth estimation, object detection, and character recognition, boosting technical proficiency across attendees

Computational Intelligence and Operations Lab (CIOL) — September 2024 – Present
Shahjalal University of Science and Technology — *Research Collaborator*

- Conducted research on hate speech detection across multilingual datasets, addressing model bias and improving classification metrics
- Designed and implemented advanced solutions for Retrieval-Augmented Generation (RAG) tasks, enabling seamless integration of external knowledge retrieval into language models and enhancing their contextual understanding, improving F1@k, MRR, precision, and recall

Projects

SpeechAya | *Multilingual LLM that can hear and speak* — August 2024 – September 2024

- Engineered a novel multilingual LLM pipeline integrating speech and text modalities, processing over 1000 hours of audio data from LibriSpeech and Mozilla CommonVoice datasets across 5 languages.
- Implemented and optimized speech tokenization using state-of-the-art models (MMS, mHuBERT, XEUS), reducing processing time by 32% through efficient batching and parallel processing.
- Achieved a score of 112 in Word Error Rate (WER) on the PolyAI/minds14 benchmark dataset by fine-tuning a Qwen2-1.5b model architecture with custom speech embeddings.
- Developed a modular training pipeline supporting multiple speech tasks (ASR, TTS, voice cloning, translation) through a unified model architecture.

Advanced Attribute Extraction and Classification Pipeline — July 2024 – August 2024

- Applied advanced OCR techniques with pre-trained models to extract text from over 400,000 product images, achieving an 88% text recognition accuracy and significantly enhancing data extraction efficiency.
- Fine-tuned DistilBERT and LLaMA 3.2 for Named Entity Recognition (NER) tasks, using proper metrics for optimization, which resulted in an improvement in entity extraction precision and recall.
- Optimized LayoutLM for attribute classification tasks, such as identifying product dimensions (e.g., weight, height, width), reducing misclassification rates (False Positives) by 10-15% and streamlining attribute extraction workflows.

Sanskriti Bench | *Benchmark for Multilingual Indic LLMs* August 2024 – Ongoing

- Contributed to the development of a novel Indian cultural benchmark, collaborating with native speakers from diverse regions across India, ensuring the dataset reflects authentic cultural nuances and linguistic diversity.
- Facilitated data collection by reaching out to elders within communities for valuable cultural insights, ensuring that all data considered for benchmarking is human-generated and contextually accurate.
- Conducted comprehensive experiments to gather relevant data for large-scale language models (LMs), designing reasoning experiments with precise metrics to enhance benchmarking accuracy and model performance.
- Pioneered synthetic data generation techniques for Hindi language processing, contributing to the creation of culturally contextualized datasets.
- Experimented with multiple language models, including LLaMA 3.3, achieving benchmark accuracies ranging from 60% to 75%, providing insights into model performance across this dataset.

Publications

[1] **Lexical Reranking of Semantic Retrieval (LeSeR) for Regulatory Question Answering :** Jebish Purbey, Drishti Sharma, Siddhant Gupta, Khawaja Murad, Siddartha Pullakhandam, Ram Mohan Rao Kadiyala *Accepted at RegNLP @ COLING 2025*

[2] **SeQwen at the Financial Misinformation Detection Challenge Task: Sequential Learning for Claim Verification and Explanation Generation in Financial Domains** Jebish Purbey, Siddhant Gupta, Nikhil Manali, Siddartha Pullakhandam, Drishti Sharma, Ashay Srivastava, Ram Mohan Rao Kadiyala *Accepted at FinNLP-FNP-LLMFinLegal @ COLING 2025*

[3] **Multilingual Hate Speech Detection and Target Identification in Devanagari-Scripted Languages** Siddhant Gupta, Siddh Singhal, Azmine Toushik Wasi *Accepted at Chipsal @ COLING 2025*

Technical Skills

Languages: Python, C++, Julia, JavaScript
Libraries and Frameworks: Pytorch, Django, Tensorflow, Sklearn, Librosa, NLTK, Trl, Transformers, LoRA, OpenCV, Numpy, Pandas, Matplotlib , Gradio, BitsandBytes
Databases: PostgresSQL, SQL , SQLite

References

Suman Debnath - Amazon
Principal Developer Advocate — suman.san14@yahoo.in— 96201 02221