Siddhant **Gupta**

Undergrad Researcher, Indian Institute of Technology Roorkee

ℹ sidworks01.github.io   @ siddhant_g@me.iitr.ac.in   github.com/SidWorks01   🎓 Google Scholar

Education

Oct 2022
July 2026

Indian Institute of Technology Roorkee [🌐] **Roorkee, India**
B.Tech student in Industrial Engineering
Coursework: Data Mining, Probability and Statistics, Calculus, C++

Experience

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

- 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
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 and improving F1@k, MRR, precision and recall.

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] **[4th position in workshop]**

[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] **[3rd position in workshop]**

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

Projects

Foundation Models for Mathematical Reasoning and Benchmarking Ongoing

Developing benchmarks and evaluation methods for mathematical reasoning

- Extracting dataset and synthetic datasets for solving complex mathematical reasoning tasks for building of foundation models, improving logical inference capabilities.
- Designing standardized benchmarks to evaluate model performance across diverse mathematical problem types.
- Experimenting with novel techniques to enhance symbolic computation and reasoning consistency in large models.

LLMs as a Judge Ongoing

Received $2000 worth of compute for experimentation

- Developing a framework for assessing the interpretability and bias in judgment outcomes across different model architectures.

Typhoon Intensity Prediction and Advanced Image Processing Ongoing

Proposing a unique and computationally efficient solution

- Designed a novel solution using traditional machine learning methods and advanced image processing techniques, achieving better time complexity compared to YOLO-based solutions.
- Collaborating on refining and publishing the model to establish its robustness in meteorological applications.

3D Tomography Image Annotation of Protein Types Ongoing

Enhancing protein structure analysis

- Implementing 3D tomography techniques to annotate protein types, aiding in structural analysis and biological research.

Multimodal Conversational AI Ongoing

Building advanced dialogue systems with multimodal inputs

- Developing a conversational AI framework capable of processing and integrating text, audio, and image modalities for seamless interactions.

SpeechAya : Speech Synthesis August 2024 - September 2024

Open-Source 8-week long Hackathon Project by Cohere4AI

- 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

Amazon ML Hackathon 2024

- Applied advanced OCR techniques with pre-trained models to extract text from over 400,000 product images, achieving a 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

Awating Submission , Ongoing Project

- 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 .
- 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.

Carbon Footprint Detector

February 2023

Full-Stack Scalable extension made in 3 days

- Created an innovative Chrome extension that analyzed carbon emissions generated by 200+ websites, resulting in a 30% increase in user engagement with sustainability metrics.
- Developed a PostgreSQL database to manage 100,000+ user records and emission metrics, integrated with a Node.js backend for real-time data analysis.
- Engineered a robust CI/CD pipeline that streamlined testing and deployment processes, resulting in an acceleration of development cycles by 30% while ensuring consistent application scalability through containerized Docker components.

DocAI

March 2024

Full-Stack Application made in 3 days

- Developed a full-stack Django web application to streamline medical test report operations, enabling seamless interactions between two distinct user roles (e.g., doctors and patients).
- Integrated NLP-based suggestion features for automated report generation, reducing manual input by 40%.
- Built a data analytics dashboard using Plotly for real-time insights and trends, and embedded a chatbot widget to assist users with suggestive use cases, improving user satisfaction by 30%.

Music Genre Classifier

April 2023 – May 2023

Audio Classification Model

- Engineered a model to classify music genres using Librosa for signal processing, achieving an 91.2% accuracy rate across a 500+ hours and 6 genres dataset of music samples.
- Enhanced a CNN model with advanced techniques such as early stopping, weight decay, dropout, and batch normalization, resulting in a 38% reduction in overfitting and boost in accuracy.
- Implemented ensemble learning methods, including bagging, boosting, and voting, to improve prediction robustness and generalization.
- Optimized hyperparameters using GridCVSearch, for a better selection of models.

Deep Space Image Classifier

March 2023

Celestial Object Classification Pipeline

- Developed a pipeline for classifying celestial objects using deep learning techniques, focusing on high-resolution image data.
- Conducted data preprocessing, augmentation, and multiclass labeling to handle imbalanced datasets effectively.
- Designed a multiclass classifier to predict black hole types, achieving 78% accuracy on astrophysical datasets.
- Conducted extensive Exploratory Data Analysis (EDA) and implemented imputation techniques such as KNN imputation and mean imputation, comparing their impact on model performance.
- Evaluated and deployed multiple algorithms, including Support Vector Machines, Random Forest Classifier, Logistic Regression, Artificial Neural Networks, LightGBM, CatBoost, and XGBoost, to ensure optimal performance.

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	PostgreSQL, SQL, SQLite

References

› Suman Debnath	*Principal Developer Advocate , Amazon, USA* [🌍]
› Jebish Purbey	*Instructor & Research Assistant, IoE, TU* [🌍]