

SIDHARTH BABU

☎ (+1) 408-649-0393 ✉ sidharthbabukish121@gmail.com 🔗 [linkedin.com/in/sidharth-babu](https://www.linkedin.com/in/sidharth-babu) 🌐 [SidharthBabu121](https://www.github.com/SidharthBabu121)

Education

University of California, Santa Cruz

Masters of Science in CS (Natural Language Processing) - GPA: 3.7

Sep. 2022 – June 2024

Santa Cruz, USA

University of California, Merced

Bachelor of Science in Computer Science and Engineering(CSE)

August 2018 – May 2022

Merced, USA

Technical Skills

Programming: Python, C/C++, Java, JavaScript, HTML/CSS, R, MATLAB, React

Frameworks: PyTorch, TensorFlow, Keras, HuggingFace, Spacy, NLTK, Pandas, OpenCV, NumPy, Matplotlib, SciPy

Developer Tools: Git, Docker, Kubernetes, Amazon Web Services (AWS), VS Code, Visual Studio, PyCharm, Eclipse

Work Experience

UC Santa Cruz, Graduate Student Researcher

Jan 2023 – Dec 2023

NLP Engineer | *Python, NLP, LLM, pBBLEU, UniEval, Sentiment Analysis*

Santa Cruz, CA

- Developed a pipeline for synthetic dialogue generation with Dr. Marilyn Walker using prompt-based transfer learning and knowledge-grounded dialogue acts across **7+** domains for Open Domain Dialogue Systems using LLMs.
- Compiled a utterance-level corpus comprising over **132k labeled utterances**, achieving a **semantic accuracy score over 77%** and a **97% dialogue act accuracy** score for generating a comprehensive dialogue-level corpus.
- Attained a **competitive UniEval scores (over 85%)** for the dialogue corpus; targeting coherence, naturalness, and engagingness to create interesting and engaging conversational interactions.

Sandia National Labs, TITANS – Math & Analytics Intern

June 2022 – August 2022

Data Scientist | *Python, Pytorch, OpenCV, Kubernetes, Pandas, Seaborn, SciPy*

Albuquerque, NM

- Conducted a detailed analysis of correlations and relationships within measured variables, revealing valuable insights into electrical test data for Sandia's commercial-off-the-shelf (COTS) program.
- Utilized and improved upon exploratory data analysis techniques to compare test data from over **8 different 3rd party Sandia affiliated sites** before generating appropriate data visualizations for new use case opportunities.

Lawrence Livermore National Labs, Summer Data Science Intern

May 2021 – June 2021

ML Engineer | *Python, Tensorflow, Keras, Pandas, Matplotlib, Astropy, OCR*

Merced, CA

- Developed and optimized several supervised and unsupervised machine learning models with a **95.4% precision and 96.3% recall rate** on recent time-domain optical astronomy data to detect, distinguish, and characterize Near Earth Objects (NEOs).
- Delivered a **25 minute presentation to 50 LLNL personal** detailing findings and potential applications of the developed predictive models and their corresponding solutions.

DOE Joint Genome Institute, UCM-JGI Summer Research Intern

June 2020 – August 2020

Data Scientist | *Python, Pytorch, AWS, SciPy*

Merced, CA

- Evaluated and enhanced a hybrid clustering strategy with the bio-python library on the Lawrence Livermore Super Cluster, resolving the big data challenges in metagenomic assembly for a dataset of **over 30,000+ reads**.
- Contributed to the testing and implementation of a new scalable algorithm, Apache SpaRC, **achieving an 80% to 90% accurate** taxonomic profile for the de novo reconstruction of over 8 artificial BioCrust genomes.

Personal Project Highlights

Deep Learning and Machine Learning Projects | *PyTorch, Python, HuggingFace, LLM*

- BabyLM** - A shared task co-sponsored by CMCL and CoNLL. The task primarily aims for small scale language modeling, human language acquisition, low-resource NLP and Cognitive modeling.
- Clickbait Spoiling via Spoiler Generation** - Implemented and strengthened several natural language processing models, achieving a **16.3% accuracy improvement** in generating spoiler sentences to combat clickbait. Built two transformer based extractive Q&A models BERT-BASE-SQuAD-2 and RoBERTa-BASE-SQuAD-2 with **over 89% in accuracy** to improve spoiler generation.

Project Protect: HealthHost | *Javascript, React, HTML/CSS, Backend Dev, Frontend Dev*

- Developed and launched a multi-platform mobile application to improve access to healthcare for the multi-ethnic communities in Merced County with over **500+ users**.
- Implemented audio features for Hmong, English, and Spanish sections while collaborating with the Content Team to update a non-coder-friendly maintenance manual to facilitate new markets and better serve the existing client base.
- Won **top honors and a \$2,500 cash prize** at the Fall 2019 Innovate to Grow Competition within Merced county.