This summer I was a Data Science Intern at Structure Therapeutics, a biotech specializing in making small oral drugs for weight loss. As a single intern, my main goals were to improve their runtime in their internal clinical monitoring app in R, which also needed to be cleaned and validated. I was able to modularize their ~10k line codebase with encapsulation and was able to decrease their backend runtime by ~80 seconds, exploring methods of caching and memoization. As a secondary project, I wanted to predict dropout in patients for one of their clinical trials. I reached out to clinical pharmacologists and clinical scientists for guidance on which features could be significant with correlating with patient retention, which was extremely helpful. Using data science techniques, I used static variables to attempt to model the data to predict dropout risk scores, but realized sequential data could also be used in combination. Using an LSTM to account for each day, I created a cascading pipeline that integrated the high recall of XGB and precision of the LSTM for an f1 score of ~72% and accuracy of ~88%, significantly helping the clinical team in potentially identifying high risk patients. Near the end of my internship, I also explored how the company could build an internal agent using an LLM foundation model with RAG. With that experience, I am motivated to apply my knowledge and learn more about fine tuning models, especially in a human-feedback context that Verifai has.

I want to be able to create an AI Agent with RAG modeling and also explore more NLP in data collection/interpretation. I also want to explore more of LangChain and how that is applied in a RAG context with the IRIS Agent conversations. I want to learn and explore as much as I can about fine tuning models, using adapters, LoRa, and even prompt engineering. Beyond the technical skills, my goal is to learn how these techniques enhance dialogue quality, improve probing strategies, and shape more productive human/AI brainstorming sessions. Ultimately, I want to leave this project with both practical implementation experience and a deeper understanding of how conversational AI can drive innovation in research.

I have significant experience with EDA and data visualization, an example being my recent summer internship, in which I helped visualize new patient metrics in development such as adverse events histograms and line plots or patient compliance rates over study days. I also independently explored clinical data for feature selection, imputation and potential data issues before building prediction models. Creating a pipeline for multiple sources of lab data, trial data, etc, to be used in a model also was important data ingestion for both the XGB and LSTM models that I created. For NLP, I've explored running BERT locally with NLTK and used latent derelict allocation in order to encode topic probability distributions for research article background descriptions. In my other previous experience at Kurma AI, I also helped fine-tune the pipeline for extracting relevant text to support RAG modeling, enabling effective encoding and retrieval of QA pairs in a vector DB.