MOTIVATION
As biomedical information in the form of publications and electronic health records (EHR) increases at an increasingly fast pace, there is clear utility in having systems that can automatically handle information extraction, summarization, and question answering tasks. While there have been significant strides in improving language tasks for general language, addressing domain-specific contexts still remains challenging. In this project, I apply and fine-tune models to the SQuAD dataset and further modify and adapt for biomedical domain-specific question answering. I evaluated and compared performance on the SQuAD dataset and BioASQ, a biomedical literature QA dataset, with the goal of analyzing and developing approaches to leverage unsupervised language models for domain-specific applications. Upon generating various fine-tuned models, the best performance for general language SQuAD QA achieved an F1 score of 76.717, EM score of 73.379, and for biomedical-specific BioASQ QA achieved an F1 score of 70.348 and EM score of 49.902.

MODEL ARCHITECTURE

RESULTS & ANALYSIS
Comparison of SQuAD and BioASQ performance

CONCLUSION
These results demonstrate that leveraging an unsupervised language model, BERT, for domain-specific QA with substantially less supervised training data achieves results that are nearing comparable to general language QA.

Karen Ouyang | kjouyang@stanford.edu
Biomedical Informatics