**BioLinkBERT for Question Answering**

**Abstract**

The main goal of this project was to implement a question answering system over an existing Synthetic Biology knowledge base. While we were able to reimplement an existing BERT based question answering model over biomedical datasets, we were not able to transfer this model to our use case in Synthetic Biology due to the absence of any Synthetic Biology question-answering datasets. In this work, we have experimented with an existing BERT based model for question answering over two biomedical datasets - BioASQ and PubmedQA. The results observed were a little different than the benchmarks in our reference paper due to slightly different hyperparameters being used but there was no appreciable improvement.

**Datasets**

Standard benchmarks for biomedical NLP tasks including question answering were based on the [BLURB](https://microsoft.github.io/BLURB/tasks.html) tasks. BLURB is a comprehensive benchmark for biomedical NLP, with 13 biomedical NLP datasets in 6 tasks. The question answering tasks include BioASQ and PubmedQA and these datasets are described in detail below.

| Dataset | Train | Dev | Test | Evaluation metric |
| --- | --- | --- | --- | --- |
| BioASQ | 450 | 50 | 500 | Accuracy |
| PubmedQA | 670 | 75 | 140 | Accuracy |

**BioASQ**

Task 10b will use benchmark datasets constructed by experts. The samples in the dataset contain a question, relevant articles and relevant snippets from that article and an exact answer.

Sample question from the dataset

Question: "Are patients with marfan syndrome at increased risk of arrhythmias?"

Context: "Marfan syndrome (MFS) is a variable, autosomal-dominant disorder of the connective tissue. In MFS serious ventricular arrhythmias and sudden cardiac death (SCD) can occur. Marfan's patients carry increased risk for cardiac arrhythmias. Ventricular arrhythmias were present in 21% and were associated with increased left ventricular size, mitral valve prolapse, and abnormalities of repolarization. Cardiac complications are rare in young patients with Marfan syndrome receiving medical therapy and close clinical follow-up. Sudden death still occurs, and appears more common in patients with a dilated left ventricle. Left ventricular dilation may predispose to alterations of repolarization and fatal ventricular arrhythmias."

Answer: "yes"

This task is divided into two phases. In Phase A, participants must retrieve relevant articles nad snippets for each question while in Phase B they must use the given snippets to get the exact answer to the question.

We only focus on Phase B for now.

**PubMedQA**

[PubMedQA](https://pubmedqa.github.io/) is a biomedical question answering (QA) dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe (e.g.: Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?) using the corresponding abstracts. Each PubMedQA instance is composed of (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

**Architecture & Experiments**

We chose [LinkBERT](https://arxiv.org/abs/2203.15827), a pretrained Language Model that leverages links between documents by training on a Document Relation Prediction task as opposed to the Next Sentence Prediction task of a standard BERT model. The authors claim that including citations allows more effective multi-hop reasoning. Specifically we use the BioLinkBERT model which was trained on the biomedical domain, using PubMed articles with citation links.

At the time of our study, BioLinkBERT had the highest score on the [BLURB](https://microsoft.github.io/BLURB/leaderboard.html) benchmark.

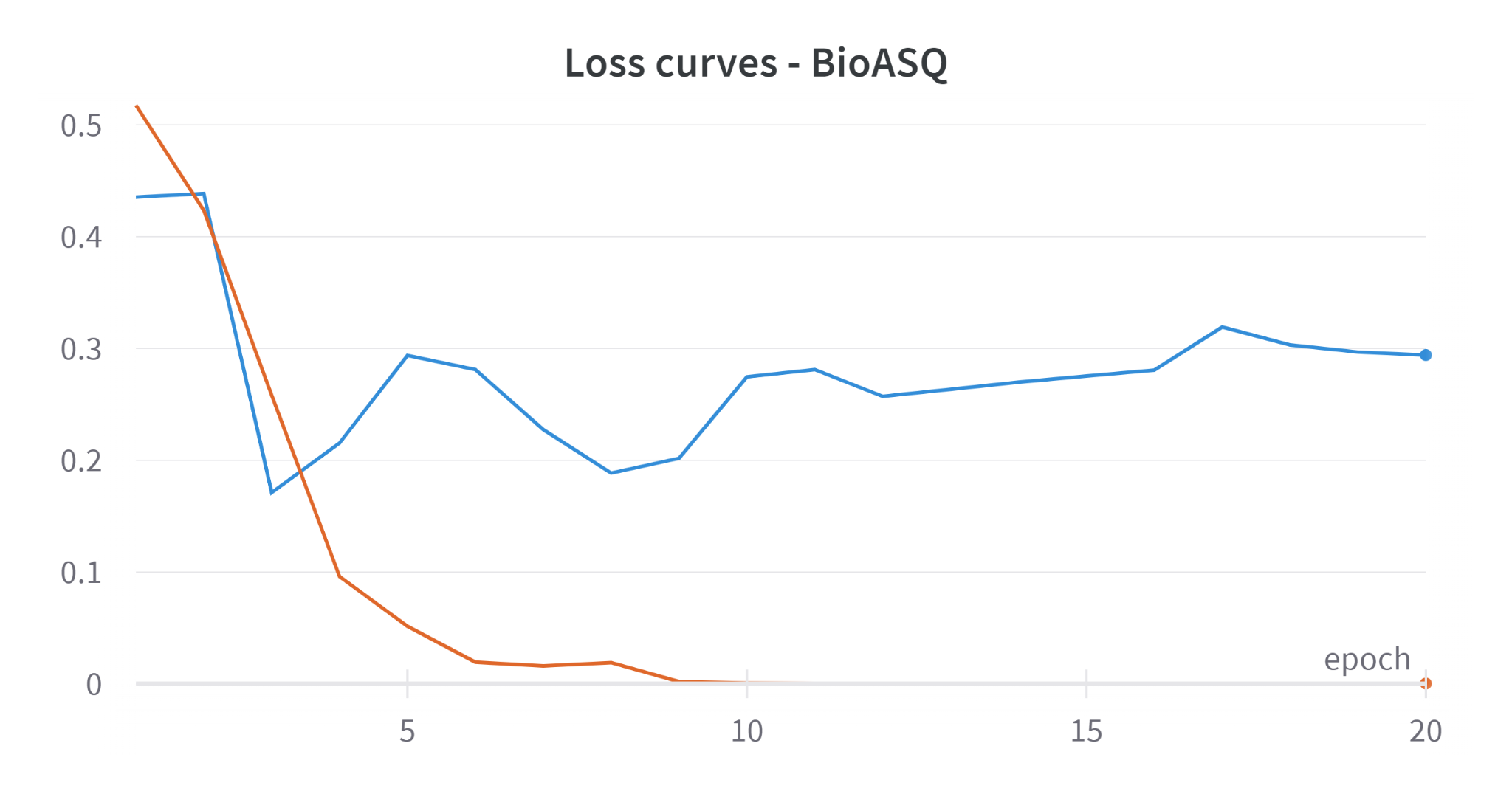
The published results in the paper are as follows

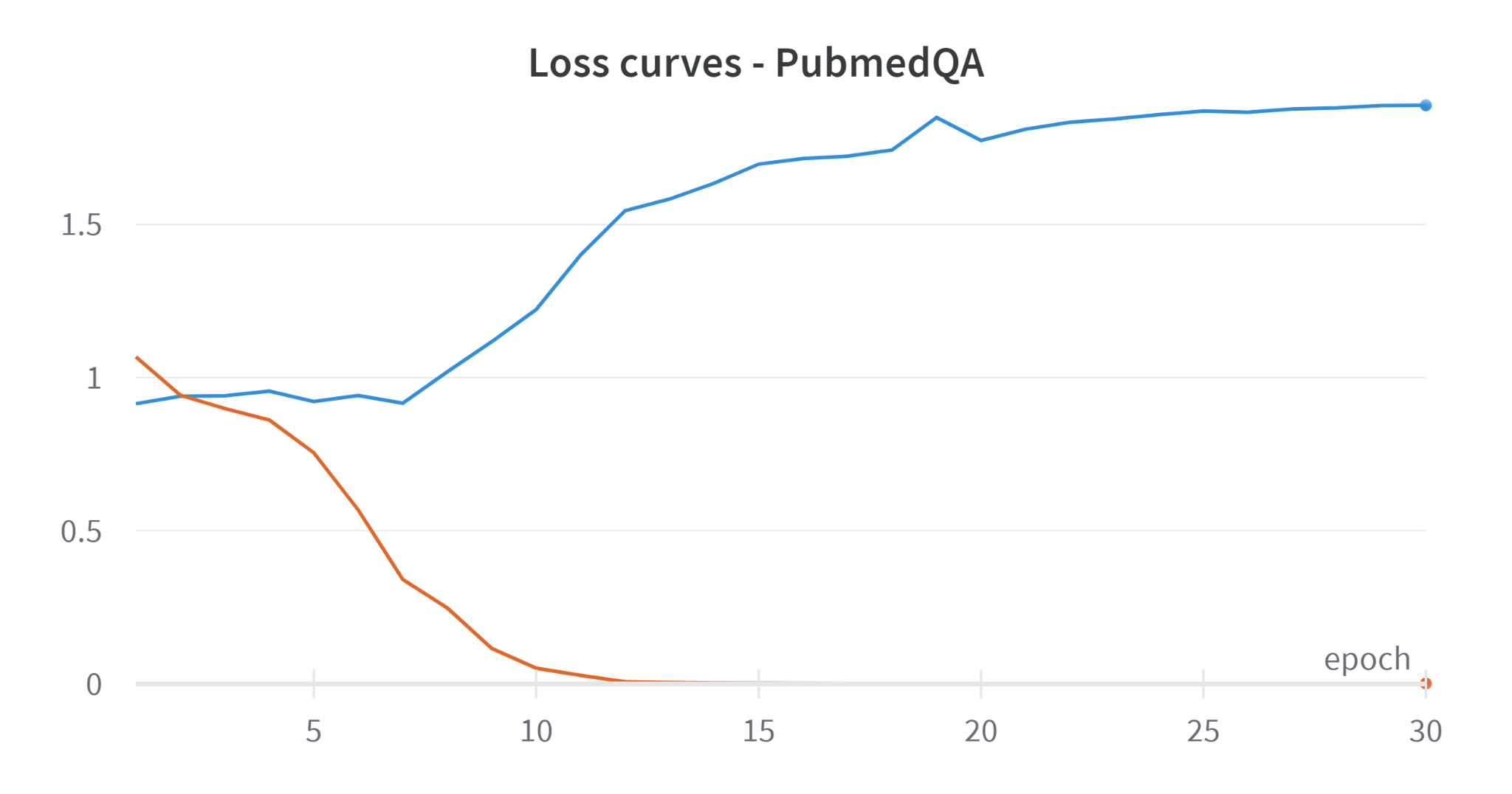
|  | PubMedQA | BioASQ |
| --- | --- | --- |
| **BiolinkBERT-base** | **70.2** | **91.4** |
| BiolinkBERT-large | 72.2 | 94.8 |

**Results**

We ran BioLinkBERT based with a training batch size of 8 instead of 16 as in the paper due to memory constraints and obtained the results below

|  | PubMedQA | BioASQ |
| --- | --- | --- |
| **Test set accuracy** | **69.2** | **95** |
| Validation set accuracy | 72 | 94.67 |





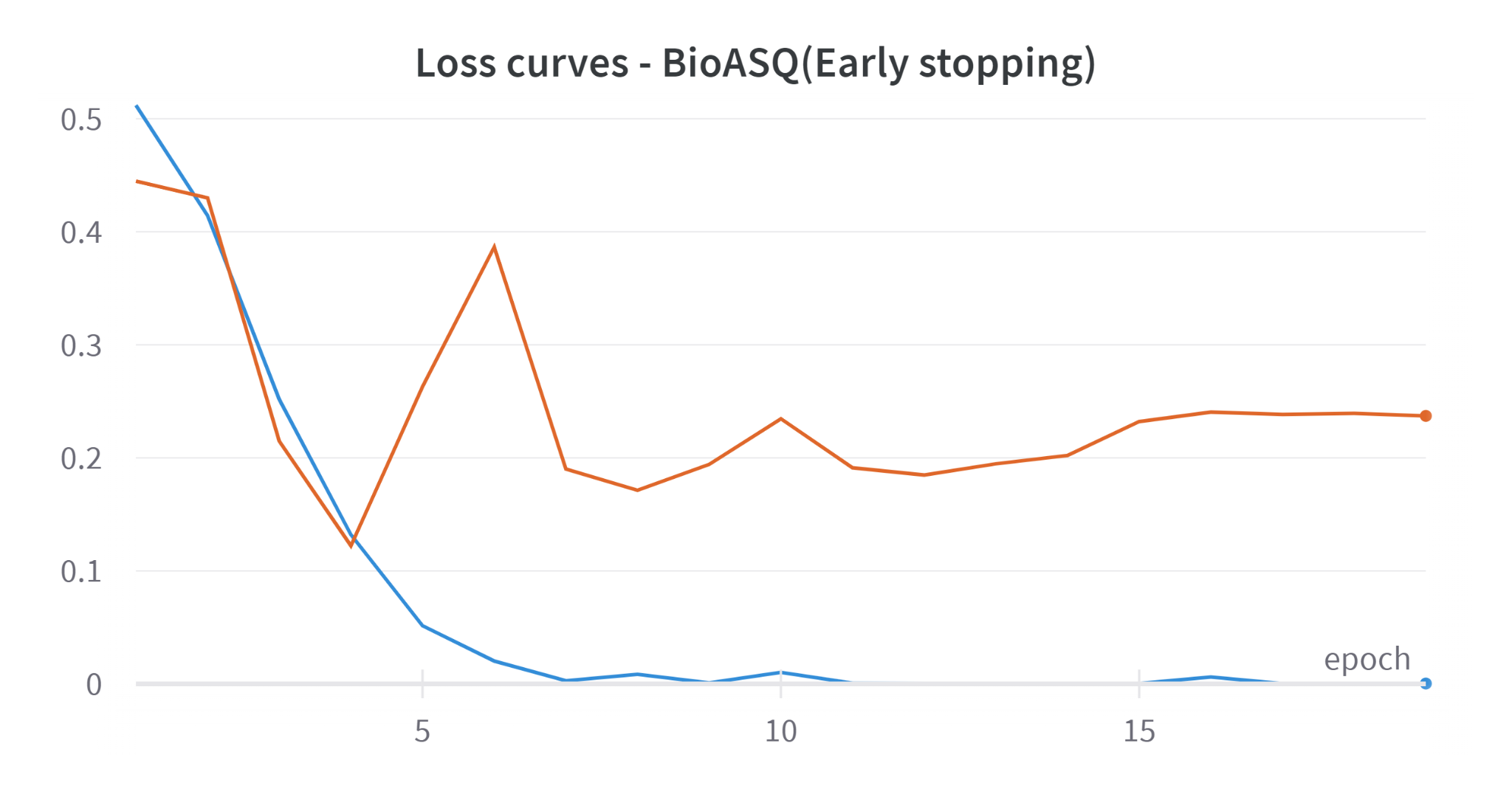
**Early stopping**

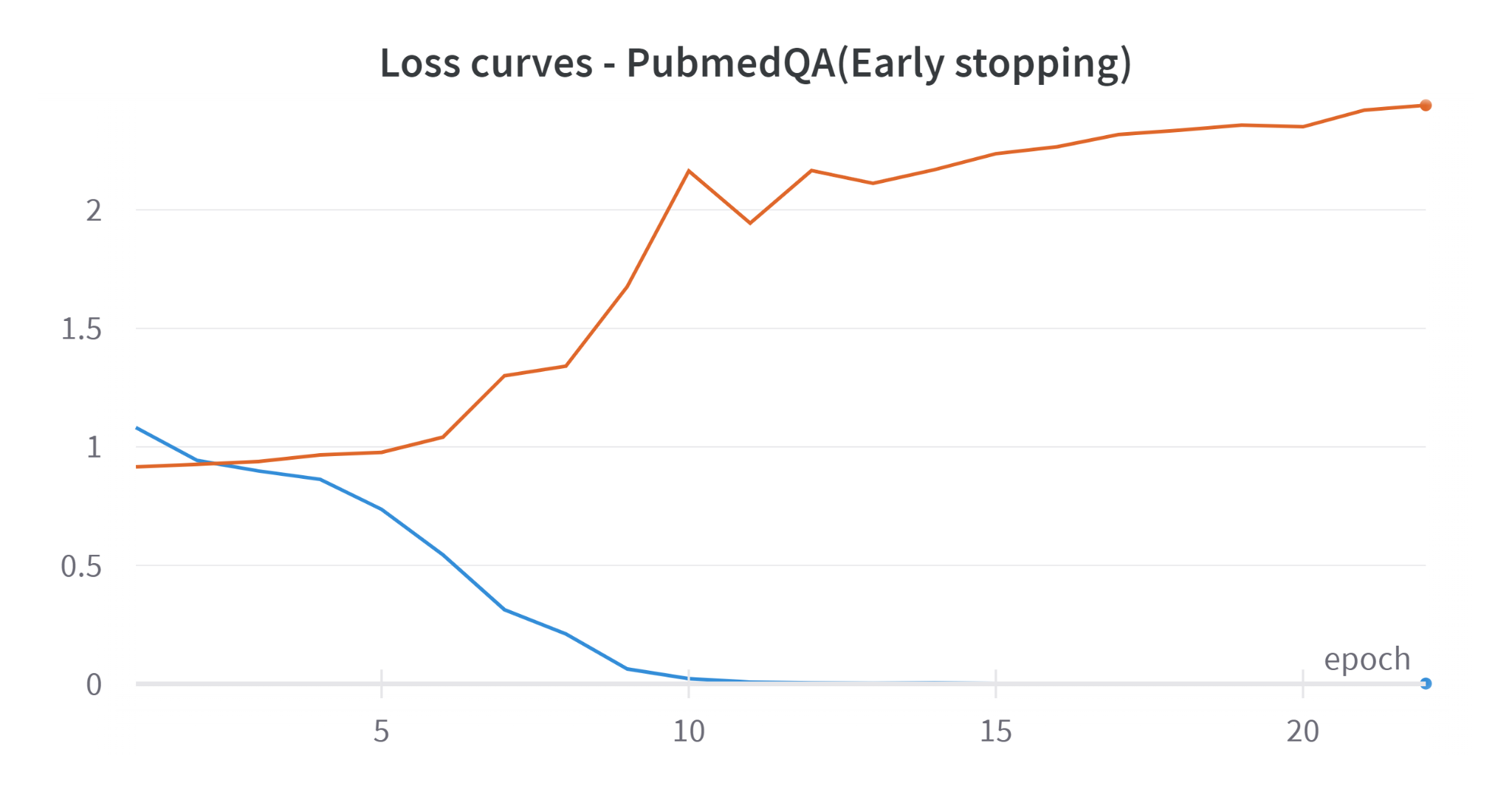
Since the loss curves for PubMedQA seem to indicate overfitting, we tried Early stopping on validation set accuracy with different Patience values. Patience value of 8 epochs gave the best results

Results with early stopping

|  | PubMedQA | BioASQ |
| --- | --- | --- |
| **Test set accuracy** | **63.8** | **93.57** |
| Validation set accuracy | 64 | 96 |

Early stopping did not really make the results better. In fact, we got slightly worse results.





**Next Steps**

1. When a similar synthetic biology question answering dataset is available in the format mentioned with a question, context text and answer, we can recreate the above work in the synthetic biology domain
2. To build an end-to-end question answering system, we must first build an Information Retrieval system that can get the relevant articles from our knowledge base and relevant snippets from the articles, based on the input question. The relevant articles can then be used as context for our current implementation. For this step, [Phase A of Task 10b](http://participants-area.bioasq.org/general_information/Task10b/) in BioASQ might be a useful resource.