# Using Out-of-Domain Contrastive Instance Bundles to Improve Medical Question Answering


## Abstract

The availability of quality, labeled data has proven to be a roadblock in the development of machine learning models for domain-specific applications. This is especially true in critical areas like healthcare, where systems that see day-to-day use must exhibit a greater standard of accuracy. There have been a number of works that examine ways to improve model performance with small data augmentation sets and training strategies, one of which being the use of perturbed instance bundles by Dua et al. in 2021. I demonstrate a strategy to improve question answering model performance on the PubMedQA dataset, a yes/no/maybe expert-labeled biomedical question answering dataset, by using contrastive BoolQ instance bundles. By pre-training models with a small set of out-of-domain bundles, I was able to improve absolute exact match and F1 score on PubMedQA by 4-7 points across all classes.


## 1 Introduction

The capabilities of large language models have dramatically improved across many fields and applications in recent years. This advancement has created new possibilities for developing AI systems that can support professionals in specialized fields where expert analysis is in short supply. Healthcare stands out as a prime example - despite requiring extensive professional staffing, many areas face persistent shortages due to the demanding educational requirements and often challenging work schedules associated with medical careers. Language models show particular promise in supporting healthcare workers by serving as knowledge assistance tools, helping to reduce the cognitive burden of memorizing extensive medical information about conditions, diseases, and treatments. This potential motivated my investigation into methods for enhancing question-answering systems specifically tailored to medical applications.

Through my research into similar work and exploration of problematic training examples, I became aware of several key challenges in developing effective architectures, datasets, and training approaches for biomedical question answering. Medical training data typically consists of lengthy questions and contexts filled with domain-specific terminology. While this level of detail and precision is crucial for medical accuracy, it can interfere with the model's ability to properly understand the semantic meaning of questions. I wanted to research whether we can enhance biomedical question-answering performance by supplementing domain-specific training data with contrastive examples that help models better grasp question semantics while avoiding the learning of irrelevant language patterns.

## 2 Similar Work

### 2.1 Learning with Instance Bundles for Reading Comprehension

Machine learning models are typically trained with the assumption that the training instances sampled from some data distribution are independent, leading to lower performance on minimally different questions. Dua et al. in 2021 introduced a method of training on instance bundles, or sets of closely contrasting examples. They were able to achieve notable improvements on HotpotQA when utilizing instance bundles from ROPES.

### 2.2 Evaluating Models' Local Decision Boundaries via Contrast Sets

Progress in NLP tasks is measured in large part by standard benchmark datasets; however, these datasets often have systematic gaps that allow simple decision rules to perform well on test data.

Gardner et al. in 2020 proposed that constructing minimally contrastive examples can characterize the correct decision boundary, leading models to perform significantly better on small input perturbations.

## 3 Methodology

### 3.1 Experiment Overview

I propose that pre-training a question-answering model on contrastive instance bundles of perturbed BoolQ questions can help dissuade learning of rules unrelated to the semantic meaning of questions and context and improve performance on the yes/no classification questions of PubMedQA. Below is a simplified figure illustrating the experiment's training pipeline.

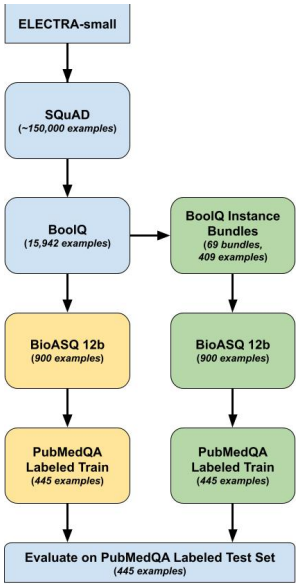


Figure 1: Experiment Training Pipeline

### 3.2 Datasets

I utilized four different datasets to train the models in my experiments, each of which are briefly detailed below.

- **Stanford Question Answering Dataset** (**SQuAD**): A reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.
- **BoolQ**: A question answering dataset for yes/no questions containing 15,942 examples. These questions are naturally occurring —they are generated in unprompted and unconstrained settings.
- **BioASQ**: A challenge for biomedical semantic indexing and question answering (QA). I specifically used the dataset for task 12b, and specifically the yes/no questions.
- **PubMedQA**: A yes/no/maybe biomedical question answering (QA) dataset collected from PubMed abstracts.

### 3.3 Experiment Setup and Base Model Finetuning

For my experiment, I selected ELECTRA-small as my base model. ELECTRA can deliver near state-of-the-art performance on question answering tasks like SQuAD 2.0 while being computationally efficient compared to other leading models. I pre-trained the model on SQuAD 2.0, achieving an Exact Match (EM) score of 78.38 and an F1 score of 86.08.

While SQuAD provides a foundation for general question answering, PubMedQA is structured differently, offering only yes/no/maybe responses. This required adding a classification layer and fine-tuning on an appropriate dataset. I narrowed the scope to yes/no questions due to the limited availability of three-class labeled data compared to binary datasets. For fine-tuning, I used BoolQ, a True/False dataset, which required simple preprocessing to convert to yes/no.

Of note: When finetuning on BoolQ, I found it very beneficial to prefix each question input with "Answer this question with yes or no: ". When evaluated on the BoolQ test set, this preprocessing results in a 9 point EM improvement.

### 3.4 Initial Analysis

Testing the initial finetuned model on PubMedQA's expert-labeled examples yielded a few interesting results. Though the model hadn't been exposed to medical domain data, its training on SQuAD and BoolQ allowed it to have performance above random chance. We can visualize its performance on the PubMedQA test set below.

These results highlight some key challenges. Without domain-specific training, the model struggled with PubMedQA content. However, the metrics reveal interesting patterns. The model showed significantly stronger performance on 'yes' responses, which could stem from multiple factors. While the dataset's slight skew toward 'yes'

| | Total | Yes | No |
|---|---|---|---|
| EM | 0.622 | | |
| F1 | 0.499 | 0.748 | 0.250 |
| Precision | 0.574 | 0.638 | 0.509 |
| Recall | 0.534 | 0.902 | 0.166 |

Table 1: Fine-Tuned Model Performance Metrics.

answers (62.31%) contributes to this bias, the dataset's size suggests additional factors are at play. A particularly notable weakness was the model's poor recall on 'no' answers, correctly identifying only 16.6% of negative responses. Analysis of false negatives revealed a pattern: questions containing positive modifiers like "Is intensive..." or "Does increased..." were often misclassified. Here's an example:

> **Question**: Does increased nerve length within the treatment volume improve trigeminal neuralgia radiosurgery?
>
> **Label**: no

Modifiers like this occur frequently in biomedical research text, and could be a large reason contributing towards the model's bias towards 'yes' predictions.

### 3.5 Contrastive Estimation

Models are typically trained to maximize the likelihood of the answer to each training question. As I am looking to maximize the benefit from introducing contrastive instance bundles, I used a combination of standard maximum likelihood estimation, and contrastive estimation (Smith and Eisner, 2005), which normalizes scores over some neighborhood of closely related instances. More specifically, I used question conditional loss, computing a probability distribution over questions in the bundle given the correct answer (Dua et al. 2021).

$$\mathcal{L}_{\text{CE-QC}}(q_g, a_g, \mathcal{B}) = \log \frac{\psi(q_g, a_g)}{\sum_{q_j \in \mathcal{B}_Q} \psi(q_j, a_g)}$$

Figure 2: Question Conditional Estimation

For contrastive instance bundles, I used a perturbed BoolQ dataset constructed by the Allen AI Institute (Gardner et al. 2020). As described in their paper, this dataset contains minimally perturbed versions of questions from the original dataset, to allow the model to learn the subtle differences in language that lead to different answers. An example is detailed below.

> **Question**: Can a tree have more than one trunk?
>
> **Label**: yes
>
> **Perturbed question**: Do some trees have more than one trunk?
>
> **Label**: yes
>
> **Perturbed question**: Do all trees have more than one trunk?
>
> **Label**: no

The dataset consists of 410 modified questions organized into 69 groups, each derived from various BoolQ questions. These modifications take different forms while maintaining the original context - such as rephrasing the question or converting it to its opposite form. This approach aims to enhance the model's understanding of both semantic relationships within questions and domain-specific content, potentially improving its ability to identify relevant text sections.

### 3.6 Biomedical Training

Finally, I trained two distinct models on domain specific data: one previously fine-tuned only on BoolQ, and another fine-tuned on both BoolQ and the instance bundles. I utilized the expert-labeled subset of the PubMedQA the data, focusing exclusively on questions with 'yes' or 'no' answers. This yielded 890 unique examples, such as:

> **Question**: Does higher body mass index contribute to worse asthma control in an urban population?
>
> **Label**: no

Since PubMedQA designates half of the labeled dataset as the test set, we only have 445 examples to train our model on. To alleviate this problem, I first trained the models on examples from the BioASQ task 12b dataset, to give the model a larger corpus of examples to form biomedical-specific connections. The dataset includes a variety of tasks, such as NER and summarization, in addition to yes/no question answering. There are 1,357 yes/no examples; however, it is heavily skewed towards 'yes' labels. To avoid label bias , I selected 1000 examples in a similar distribution of labels to the labeled subset (60% 'yes, 40% 'no'). Another potential issue with training on the BioASQ data is that the contexts given are

long: often over 512 tokens. In order to reduce the instances where answers are missed due to truncation, I implemented a document stride approach, splitting the text into overlapping chunks and averaging predictions over the context document for each QA example. After training on BioASQ data, the models are finally trained on the 445 example PubMedQA training set.

# 4 Results

## 4.1 Model Performance Metrics

The evaluation metrics of the two models when tested on the 445 example PubMedQA test set are visualized below:

| | Total | Yes | No |
|---|---|---|---|
| EM | 0.643 | | |
| F1 | 0.623 | 0.708 | 0.539 |
| Precision | 0.623 | 0.717 | 0.528 |
| Recall | 0.625 | 0.699 | 0.550 |

Table 2: Control model (not trained on bundles).

| | Total | Yes | No |
|---|---|---|---|
| EM | 0.690 | | |
| F1 | 0.669 | 0.753 | 0.584 |
| Precision | 0.670 | 0.745 | 0.595 |
| Recall | 0.667 | 0.761 | 0.574 |

Table 3: Experiment model (trained on bundles).

| | Total | Yes | No |
|---|---|---|---|
| EM | 0.047 | | |
| F1 | 0.045 | 0.044 | 0.045 |
| Precision | 0.047 | 0.027 | 0.067 |
| Recall | 0.043 | 0.062 | 0.024 |

Table 4: Difference (experiment - control).

The results show that pre-training on the BoolQ contrast set led to significant improvements across all metrics when applied to the PubMedQA dataset. Given that the contrast dataset maintains an approximately equal distribution of yes/no labels and comprises only a small portion of the training data (2.5% of BoolQ examples), these improvements likely stem from something more fundamental than simply learning label distribution patterns. The evidence suggests that exposure to the contrastive bundles enhanced the model's grasp of semantic relationships within questions and contexts, substantially boosting its question-answering performance on PubMedQA.

## 4.2 Example Analysis

To better understand how the contrast set influenced the model's performance, I analyzed which examples were correctly classified by the contrast model but missed by the control model. This analysis revealed several key patterns. First, the contrast model showed marked improvement in handling examples that included positive modifiers with negative conclusions. This is true for modifiers in the question ("Does increasing..."), as well as modifiers in the context ("statistically significant"), which might indicate positive results. This is demonstrated by the following example:

> **Question**: Does increasing blood pH stimulate protein synthesis in dialysis patients?
>
> **Label**: no

Another notable improvement appeared in questions about study results, particularly in cases where studies showed mixed outcomes. The contrast model better distinguished between positive findings in some areas but negative results in the specific area being questioned, or vice versa, as well as dealing with questions that necessitated the comparison of two different study groups. An example of such as question can be seen below:

> **Question**: Is combined therapy more effective than growth hormone or hyperbaric oxygen alone in the healing of left ischemic and non-ischemic colonic anastomoses?
>
> **Label**: yes

Overall, it seems like the addition of the perturbed bundles into the training pipeline assisted the model in a number of ways. It appeared to especially help to avoid learning common language patterns that were unrelated to question or context meaning, such as associating positive words with positive results.

# 5 Conclusion

This experiment sought to investigate whether out-of-domain contrastive instance bundles could enhance model performance on PubMedQA's medical questions. The training process incorporated four datasets, in addition to the bundles: SQuAD (150,000 examples), BoolQ (15,942 examples), BioASQ task 12b (1000 examples), and a yes/no

subset of PubMedQA (890 examples). The results demonstrated that including a relatively small contrast set (410 examples in 69 instance bundles) during pre-training significantly improved performance on PubMedQA. Despite its small size compared to other training datasets, the contrast set enhanced the model's reasoning capabilities across multiple different categories of examples, improving performance metrics on the PubMedQA test set across the board. As my experiment set out to primarily investigate the effect of the instance bundles, overall model performance can certainly be improved though more sophisticated architecture, data augmentation, and preprocessing strategies, which potentially could shed more light on the efficacy of the techniques I have demonstrated. Even with those caveats, the findings of the experiment suggest that domain-specific language models can be improved through the use of targeted instance bundles that are unrelated to the learning of technical and field specific language patterns.

## References


1. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. Retrieved from https://arxiv.org/abs/1909.06146.

2. Dua, D., Dasigi, P., Singh, S., Gardner, M. (2021). Learning with Instance Bundles for Reading Comprehension. Retrieved from https://arxiv.org/abs/2104.08735.

3. Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., Singh, S., Smith, N. A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., Zhou, B. (2020). Evaluating Models' Local Decision Boundaries via Contrast Sets. Retrieved from https://arxiv.org/abs/2004.02709.

4. Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., Toutanova, K. (2019). BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. Retrieved from https://arxiv.org/abs/1905.10044.

5. Smith, N. A., Eisner, J. (2005). Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In K. Knight, H. T. Ng, K. Oflazer (Eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 354–362). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P05-1044. DOI: 10.3115/1219840.1219884.