SMU Data Science Review

Volume 7 | Number 1 Article 9

Question Answering with distilled BERT models: A case study for Biomedical Data

Brittany Lewandowski
Southern Methodist University, blewandowski@mail.smu.edu

Rayon Morris
Southern Methodist University, rayonm@mail.smu.edu

Pearly Merin Paul
Southern Methodist University, bpr.smu2023@gmail.com

Robert Slater
Southern Methodist University, rslater@mail.smu.edu

Recommended Citation

Lewandowski, Brittany; Morris, Rayon; Paul, Pearly Merin; and Slater, Robert () "Question Answering with distilled BERT models: A case study for Biomedical Data," *SMU Data Science Review*: Vol. 7: No. 1, Article 9.
Available at: https://scholar.smu.edu/datasciencereview/vol7/iss1/9

Question Answering with distilled BERT models: A case study for Biomedical Data

Brittany Lewandowski[1], Pearly Paul[2], Rayon Morris[3], Robert Slater[4]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] 26425 Boaz Lane Dallas TX 75205 US,
Dallas, TX 75275 USA
{blewandowski, ppaul, rayonm, rslater}@smu.edu

Abstract. In the healthcare industry today, 80% of data is unstructured (Razzak et al., 2019). The challenge this imposes on healthcare providers is that they rely on unstructured data to inform their decision-making. Although Electronic Health Records (EHRs) exist to integrate patient data, healthcare providers are still challenged with searching for information and answers contained within unstructured data. Prior NLP and Deep Learning research has shown that these methods can improve information extraction on unstructured medical documents. This research expands upon those studies by developing a Question Answering system using distilled BERT models. Healthcare providers can use this system on their local computers to search for and receive answers to specific questions about patients. This paper's best TinyBERT and TinyBioBERT models had Mean Reciprocal Rank (MRRs) of 0.522 and 0.284 respectively. Based on these findings this paper concludes that TinyBERT performed better than TinyBioBERT on BioASQ task 9b data.

1 Introduction

Unstructured data imposes challenges for every industry. Due to its lack of standardization, unstructured data remains difficult to access, translate, and extract insights from (Joshi 2016). While unstructured data impacts every industry, the medical industry is strongly affected by it. Given that 80% of healthcare data is unstructured (Razzak et al., 2019), a solution is needed to help healthcare providers quickly access this data and find answers within it.

In the medical industry, unstructured data that is leveraged regularly by healthcare providers include clinical notes, "progress notes, discharge summaries," and electronic health records (EHRs) (1, Pampari et al., 2018). While EHRs provide comprehensive patient data, the challenge with such large amounts of data is that it can be difficult for healthcare providers to access it efficiently. Today, healthcare providers must search through patient record tabs to find information they are looking for. This process is time-consuming and takes healthcare providers away from caring for their patients.

To emphasize how time consuming it is for healthcare providers to search for answers in EHRs, Overhage et al. (2020) conducted a research study that found that healthcare providers spent 16 minutes per EHR reviewing its contents. With an assumption that healthcare providers see 20 patients a day, this means that 5 hours of their workday is spent reviewing EHR documents. Another study conducted by Khairat et al. (2021) surveyed 25 healthcare providers to determine how difficult it is for these individuals to extract relevant information from EHRs. Results of this study showed that most of the surveyed healthcare providers indicated that it was difficult finding information that they needed to care for their patients. Although EHRs are valued medical records, the two studies above illustrate the need for a solution that can help healthcare providers efficiently access patient data.

A solution to overcoming the challenge of efficiently extracting information from EHRs is machine learning. Specifically, researchers have found that leveraging Natural Language Processing (NLP) and Deep Learning are effective at standardizing unstructured data. Together, these methodologies transform text-heavy data assets and develop models that can be used for Tasks such as Question Answering (Hugging Face 2022).

Question Answering (QA) is a technique within NLP and Information Retrieval (IR) that provides natural language answers to questions asked by users (Sarkar, 2016). In the context of this research, a Question Answering system could be leveraged by physicians to decrease the time they spend reviewing EHRs, by allowing them to ask questions to the system about patient data. Historically, Question Answering has been applied on generalized, open domain use cases. However, Question Answering has been used less frequently on closed domain use cases such as Electronic Health Records (EHRs).

Considering other research that has been done in this space, previous researchers have used distilled Bidirectional Encoder Representations from Transformers (BERT) models to develop Named Entity Recognition models (Zöllner et al., 2021) and Question Answering systems using an ensemble approach (Wu et al., 2021). However, there is limited research on using distilled models alone to build Question Answering systems. Consequently, this research aims to build on this foundation by using TinyBERT and TinyBioBERT models to create a closed domain Question Answering system. Unstructured data contains a wealth of information that healthcare providers can leverage to aid their decision making. This research aims to develop a solution to achieve this and enhance healthcare providers' productivity.

2 Literature Review

2.1 Question Answering Systems

Building a deep learning application that can successfully answer questions on a determined subject remains a long-standing problem (Widad et al., 2022) within Natural Language Processing (NLP) and Machine Learning (ML). To improve upon model results, research has proposed datasets on which these NLP based models can be tested. Several examples of these datasets include Stanford Question Answering Dataset (SQuAD), and MS Marco (Widad et al., 2022). Using both a BERT model and the SQuAD dataset, (Widad et al., 2022) achieved results exceeding 70% for their COVID-19 based Question Answering system. Another example of successfully building a Question Answering system on a use case specific data set is shown with (Alzubi et al., 2021). In their research, the authors used COBERT, a healthcare application of BERT, on a dataset made available by the Coronavirus Open Research Dataset Challenge that achieved an Exact Match (EM)/F1 score of 80.6/87.3 (Alzubi et al., 2021).

Although Question Answering systems built on pre-defined data sets increases model accuracy, computers remain challenged in understanding and extracting questions that are asked to it. As a solution to this challenge, research has proposed using RQE, or Recognizing Question Entailment for Question Answering. RQE “retrieves answers to a premise question (PQ) by retrieving inferred or entailed questions, called hypothesis questions (HQ) that already have associated answers" (1, Abacha et al., 2019). RQE differentiates itself from other Question Answering approaches by using a "multi-step approach that tackles the challenging issues of question understanding and answer extraction in a unique way" (2, Abacha et al., 2019). Using the evaluation process used in TREC 2017 LiveQA, (Abacha et al., 2019) achieved an improvement of 29.8% over the best official score (IR+RQE System 0.827, LiveQA'17 Best Result 0.637).

Question Answering systems have made significant progress since their inception in the 1960s. Although many models have been developed since then, this area of research continues to evolve as new methodologies and models become available that increase performance with fewer parameters. This research highlights the progression of Question Answering systems by building a system on the BioASQ dataset using distilled models. This research is cognizant of the fact that as models become increasingly distilled, it can have an adverse effect upon model results due to information loss (Gao et al., 2020).

2.2 BioASQ

Recognizing that NLP based models can perform well with domain specific data, BioASQ has developed a series of biomedical data sets that can be leveraged in machine learning models to improve access to unstructured text data. In addition to providing the public with these data sets, the organization has also been organizing community-based “challenges” where teams can compete against one another to build the best performing model on a defined BioASQ data set. When this paper was written, several active challenges included disease indexation and Biomedical question answering.

To provide an overview of a typical BioASQ challenge, (Nentidis et al., 2021) published a paper highlighting their participation in a 2019 challenge. For their challenge, (Task 7), teams had the choice of competing in Task A, Task B, or both. Task A focused on semantic indexing, while Task B focused on developing a Question Answering system. After deciding what Task each team wanted to pursue, they began developing their systems and submitting their results. For Task 7b, Nentidis et al. also discussed the two phases that teams could participate in (Phase A and Phase B). In Phase A, participating teams were instructed to extract and surface the most relevant information to answer a question, while Phase B instructed teams to surface the proper answer for a given set of questions. In both Task A and Task B, 30 teams participated. Relevant to Task 7b, developing a question answering system, 13 teams participated, all of which developed their systems using models including: BERT, BioBERT, ELMo, and BiLSTM (Nentidis et al., 2021).

2.3 Transformer Models

As the amount of unstructured data continues to increase, research has focused on using NLP-based transformer models to extract useful information from text-based documents. As opposed to Recurrent Neural Networks (RNNs), transformer models utilize a self-attention mechanism that enables them to understand word contexts regardless of where words reside in text. Details on how Transformer models work can be found in the research paper “Attention Is All You Need” (Vaswani et al., 2017). Additionally, transformer models can process long input data in one pass, solving many of the challenges associated with natural language models (Chaves et al., 2022). In this paper, all models used to build the Question Answering system are transformer based, which allows for faster model training time and improved model performance (Chaves et al., 2022).

2.4 BERT & Distilled BERT Models

In the research community, transformer models such as BERT have become popular for standardizing unstructured text data. BERT distinguishes itself from other transformer models by leveraging a bidirectional approach to training. This bidirectional approach enables BERT to understand the context surrounding text corpora, making it a more powerful NLP model (Devlin et al., 2018). The challenge with BERT, however, is that it is compute-intensive (Zhao et al., 2021). $\text{BERT}_{\text{BASE}}$ uses 12 layers and 110 million parameters (Hugging Face 2022), which does not scale well on resource-constrained devices. This means running BERT models on local computers is difficult to accomplish (Xiaoqi et al., 2019).

Since BERT is regarded as one of the best transformer models in the market, distilled versions of the model have been developed as a workaround. Several of these models include TinyBERT and TinyBioBERT. Research on these distilled models has shown that they are effective alternatives to the $\text{BERT}_{\text{BASE}}$ model, consuming less memory and producing comparable evaluation metrics. Research has shown that TinyBERT consumes 1.2 times less memory than $\text{BERT}_{\text{BASE}}$ when using the same batch size (Rohanian et al., 2022).

Compared to $\text{BERT}_{\text{BASE}}$, TinyBERT is 7.5 times smaller with four layers (Xiaoqi et al., 2019), and TinyBioBERT is 15 times smaller with four layers and 15 million parameters (Mahdinoori, 2022). By reducing the number of layers and parameters used in these distilled models, BERT models are accessible on local computers. This is a significant achievement, as insights gained from BERT models can be distributed to anyone who accesses a computer for their job. For healthcare providers in the medical industry, this means gaining more access to patient information and reducing the time spent searching for answers buried in unstructured data.

2.5 BioASQ for Question Answering Systems

Since BioASQ contains large biomedical Question Answering data sets, many research papers have been published using BioASQ data to develop Question Answering systems for the medical industry. The hope is that by developing Question Answering systems with biomedical-specific data, these systems will be transferrable to medical data used in the field.

One example of research that used BioASQ data to build a Question Answering system is Sarrouti et al. in which the authors built a Question Answering system based on a multi-class SVM model and handcrafted lexico-syntactic patterns. This model achieved an accuracy of 89.4%, proving its effectiveness at surfacing answers for users (Sarrouti et al., 2020).

Another research paper that developed a Question Answering system using BioASQ data was Xu et al. in which the authors built a Question Answering system on BioASQ tasks 6b, 7b and 8b. Unique to this study was the approach to building the system. Upon recognizing that biomedical documents are highly specialized and possess unique structures and entities, the authors implemented a specialized feature engineering framework that included part of speech (POS) tagging, and NER (Name Entity Recognition) to ensure that generalizable features of text documents were identified (Xu et al., 2021). The idea is that by implementing this framework, Question Answering systems will be better poised to accurately identify and answer user questions. Using this framework, the authors built their system using BioBERT and saw an improvement in their Strict Accuracy (SAcc), Lenient Accuracy (LAcc), and Mean Reciprocal Rank (MRR) as shown in Table 1.

Table 1. Xu et al. BioASQ Results.

BioASQ Task:	SAcc, LAcc, & MRR Scores:
6B	(0.4517, 0.6294, 0.5197)
7B	(0.444, 0.6419, 0.5165)
8B	(0.3937, 0.6098, 0.4688)

Other research includes using deep learning for Question Answering systems. Deep learning algorithms are highly effective; however, they require substantial amounts of training data which is not always accessible in the medical industry. Transfer learning has been employed to apply deep learning algorithms on small data sets to overcome this obstacle. Specifically, Du et al. used a combination of BERT, LTSM, ATT & GRU to build their question answer dataset. In this research, their model was trained on BioASQ task 5b data and showed that neural Question Answering systems compete with other non-neural systems producing an MRR of 0.7 (Du et al., 2020).

2.6 Other Question/Answering Research:

Researchers have leveraged data sets, including PubMedQA, to build Question Answering systems. One such study looked to enhance Question Answering systems by building a sentiment-aware learning pipeline (Zhu et al., 2022). The pipeline used BioBERT, T5, ROBERTa, and XGBoost and produced an accuracy of 83.1 and an F1 score of 76.92. These metrics were improvements from previous models that used BioBERT and multi-phase BioBERT alone.

Other research has built Question Answer systems using scholarly scientific articles with Question Answer pairs. Specifically, one study used 100,000 human-annotated context-Question Answer triplets to build a Question Answering model. The research of Tanik et al. fit their data to BERT, SciBERT, and Bi-DAF models, and found that SciBERT performed the best, F1 score 75.46. (Tanik et al., 2022).

3. Methods

3.1 Data

Data used in this research comes from task 9b of the BioASQ competition. Provided with Task 9b is a benchmark dataset, (training set), and five gold standard datasets, (test sets), that were developed by biomedical experts. Both datasets are JavaScript Object Notation (JSON) files that organize biomedical questions in a series of nested dictionaries. Table 2 below shows the train and test set formats, and Table 3 provides an example of a question's format.

Table 2. Train & Test Set Data Formats.

Keys In Datasets:	Information Provided by Key:
type	Identifies what category a given question falls into (factoid, yes/no, summary, or list).
body	Contains the biomedical question being asked.
id	Unique identifier for a question.
ideal_answer	Paragraph for a given question.
exact_answer	Exact answer for a given question.
documents	Link to a PubMed article where the answer to a given question can be found.
snippets	Section of PubMed article where the answer to a given question can be found.
concepts	Entities or attributes relevant to a given question.
triples	Contextualizes questions by identifying their subject(s), predicate(s) and object(s) in each question.

Table 3. Example of Data Provided for A Factoid Question.

Keys In Datasets:	Sample Value for A Given Key:
documents	http://www.ncbi.nlm.nih.gov/pubmed/33186545
text	De Novo VPS4A Mutations Cause Multisystem Disease with Abnormal Neurodevelopment.
type	factoid
body	Which disease is caused by de novo VPS4A mutations?
id	601bde6e1cb411341a000006
ideal_answer	De vovo VPS4A mutations cause multisystem disease with abnormal neurodevelopment.
exact_answer	Multisystem disease with abnormal neurodevelopment

3.2 Data Pre-Processing

For this research, pre-processing was done to both the benchmark, (training), and golden, (test), data sets. Before pre-processing, the BioASQ training dataset contained questions organized into the following four categories: Yes/No, Factoid, List, and Summary. Since Question Answering systems are context dependent and require validation, this research filtered the training dataset to only include factoid questions.

Next, the five batches of golden, (test), data sets were pre-processed. Given that the golden data sets contained answers in them, it was recognized that they would produce overfit models. Consequently, this research used a script to strip the answers out of the golden datasets.[1] Once the answers were removed, this research saved the output into five new files that were used in model evaluation as test datasets.

Finally, recognizing that large training sets help transformer models achieve more favorable results, pre-processing was done to increase the number of questions contained in the training dataset. In the BioASQ datasets, one question can be paired with multiple text

[1] https://github.com/urvashikhanna/bioasq9b/blob/main/transform_n2b_factoid.py

snippets from PubMed articles. Consequently, this research increased the size of the train dataset by running the BioASQ evaluation script that assigned one text snippet to every factoid question. To avoid duplication of question IDs, the script appended three numbers to the end of question IDs that had multiple text snippets, ensuring that they remained unique. Originally, one question may have been paired with multiple text snippets; however, after pre-processing, every question in the train and test sets was paired with one text snippet. An example of what a given factoid question looked like before and after pre-processing can be found in Figures 1 and 2 below.

documents: http://www.ncbi.nlm.nih.gov/pubmed/33186545
id: 601bde6e1cb411341a000006
type: factoid
body: Which disease is caused by de novo VPS4A mutations?
text: De Novo VPS4A Mutations Cause Multisystem Disease with Abnormal Neurodevelopment.
document: http://www.ncbi.nlm.nih.gov/pubmed/33186545
text: The endosomal sorting complexes required for transport (ESCRTs) are essential for...
document: http://www.ncbi.nlm.nih.gov/pubmed/33186545
exact_answer: Multisystem disease with abnormal neurodevelopment

Figure 1. Factoid Question Before Pre-Processing.

id: 601bde6e1cb411341a000006_001
question: Which disease is caused by de novo VPS4A mutations?
exact_answer: Multisystem disease with abnormal neurodevelopment
context: De Novo VPS4A Mutations Cause Multisystem Disease with Abnormal Neurodevelopment.

id: 601bde6e1cb411341a000006_002
question: Which disease is caused by de novo VPS4A mutations?
exact_answer: Multisystem disease with abnormal neurodevelopment
context: The endosomal sorting complexes required for transport (ESCRTs) are essential for...

Figure 2. Factoid Question After Pre-Processing.

To illustrate the expansion of After pre-processing, counts of the number of questions contained in each data set was calculated. The results of these counts are found in Table 4 below.

Table 4. Count of Questions Before & After Pre-Processing.

Dataset:	Count Of Questions Before Pre-Processing:	Count Of Questions After Pre-Processing:
Training Dataset	1092	5727
Batch 1 Golden Test Dataset	29	411
Batch 2 Golden Test Dataset	34	459
Batch 3 Golden Test Dataset	36	394
Batch 4 Golden Test Dataset	28	325
Batch 5 Golden Test Dataset	37	445

3.2 Methods

This paper leverages the methodologies of machine learning and natural language processing (NLP) to create a Question Answering system for healthcare providers. Machine learning was selected as an appropriate method given that this research develops a model that extracts precise answers from context. Additionally, since this research includes unstructured text data, NLP was chosen to convert this data into a format that can be used by machine learning models.

3.3 Models

Since this research focuses on developing a Question Answering system that can be used by healthcare providers in hospitals, distilled BERT models were used. BERT was selected as an appropriate model as it is known as one of the best NLP models in the market. However, since BERT is computationally intensive, this paper uses distilled versions of it, namely TinyBERT, and TinyBioBERT. Tables 5-7 highlight the architectural differences between BERT$_{\text{BASE}}$, TinyBERT, and TinyBioBERT.

Table 5. BERT$_{\text{BASE}}$ Architecture.

Number of Layers:	12

Number of Dimensions:	768
Attention Heads:	12
Parameters:	110 million

Table 6. TinyBioBERT Architecture.

Number of Layers:	4
Number of Dimensions:	768
Attention Heads:	12
Parameters:	15 million

Table 7. TinyBERT Architecture.

Number of Layers:	6
Number of Dimensions:	768
Attention Heads:	12
Parameters:	14.5 million

Aside from the computational advantages of TinyBERT and TinyBioBERT, these models were also selected based on the data on which they are trained. While TinyBERT is trained on a distilled portion of the English Wikipedia and Brown corpus (Anderson, 2019), TinyBioBERT is trained on biomedical specific data. The intent of using both models was to compare their output and determine whether using a generalized model, (TinyBERT), or a model specific to the medical industry (TinyBioBERT), is more effective at building Question Answering systems for healthcare providers.

Although BERT, TinyBERT and TinyBioBERT differ in terms of their architectural complexity, they all use the encoder layer of the Transformer architecture depicted in Figure 3. Using this encoder layer, BERT models can process sequential data with high precision. Unlike Recurrent Neural Networks, (RNNs), BERT does not have recurrent or convolutional layers. Without these layers, these models can process data in parallel and expedite model training.

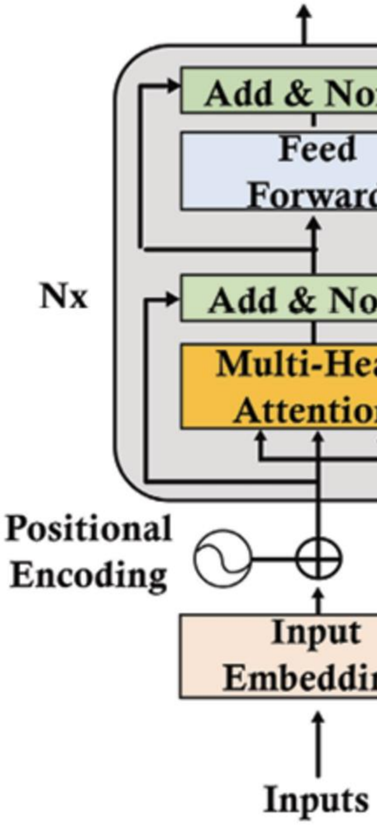

Figure 3. Architecture of BERT’s encoder layer.

Another unique component of BERT models is their multi-head attention layers. Previously, RNNs were challenged with identifying long range dependencies in word sequences. With multi-head attention layers, this challenge is resolved. In the Question Answering domain, multi-head attention layers enable the models to learn different linguistical features in word sequences that are important for producing accurate predictions. These layers operate similarly to filters in Convolutional Neural Networks

(CNNs). From a computational perspective, multi-head attention layers ingest a given input sequence's value, key, and query, and pass them through three linear layers. Next, they calculate the dot product of the key and query tensors and scale them. Finally, the scaled dot products are concatenated and passed through a final linear layer to achieve the expected output. To understand how these multi-head attention layers are calculated, Figure 4 defines the formula. Figure 5 provides a visual representation of the multi-head attention architecture.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Figure 4. Multi-Head Attention Layer Formula.

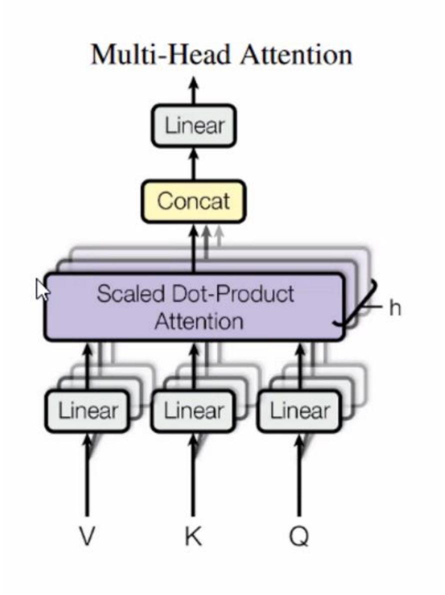

Figure 5. Architecture for Multi-Head Attention Layers.

3.4 Evaluation

Strict Accuracy (SAcc), Lenient Accuracy (LAcc), and Mean Reciprocal Rank (MRR) were the evaluation metrics used for both models in this research.[2] These metrics were selected because they are common evaluation metrics for Question Answering systems and because BioASQ's task 9b competition uses them to rank competitors' results. Additionally, given that BioASQ evaluates factoid-based Question Answering systems based off a list of the top five most probable answers in descending order, these metrics were deemed appropriate.

Relevant to the BioASQ competition, SAcc evaluates a model's ability to correctly identify the exact answer to a question in the first position of its predicted answer list. For example, if the question provided to the Question Answering system was "What color is the sky?" and the resulting predicted answer list was [blue, red, and pink] the SAcc would be 100%. This is because blue was in the first position of the answers returned. In contrast to SAcc, LAcc is "lenient" as it deems a predicted answer as correct if the exact answer is contained within the returned list. Finally, MRR evaluates the model's ability to correctly identify the exact answers by taking the average of the reciprocal rank for each question. If the answer list does not contain the exact answer, it receives a MRR of 0. A high MRR score indicates that a model is accurately ranking its results, whereas a low MRR score indicates that a model is not accurately ranking its results. Table 8 below provides the formulas used for calculating SAcc, LAcc, and MRR.

Table 8. Mathematical Formulas for SAcc, LAcc, and MRR.

Evaluation Metric:	Formula:
SAcc	$\frac{c1}{n}$
LAcc	$\frac{c5}{n}$
MRR	$\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{r(i)}$

4 Results

[2] github.com/BioASQ/Evaluation-Measures

This research fit TinyBioBERT and TinyBERT models for all five batches of BioASQ's task 9b golden test datasets. All models were fitted using a Tesla T4 Nvidia GPU with a maximum memory usage of 15109MiB. Results obtained for each test set are displayed in Tables 9 and 10 below.

Table 9. TinyBioBERT Model Results.

Batch Evaluated:	SAcc Score:	LAcc Score:	MRR Score:
Batch 1 Golden Test Dataset	0.06	0.275	0.131
Batch 2 Golden Test Dataset	0.147	0.441	0.213
Batch 3 Golden Test Dataset	0.25	0.333	0.271
Batch 4 Golden Test Dataset	0.25	0.464	0.284
Batch 5 Golden Test Dataset	0.194	0.444	0.262

Table 10. TinyBERT Model Results.

Batch Evaluated:	SAcc Score:	LAcc Score:	MRR Score:
Batch 1 Golden Test Dataset	0.206	0.517	0.278
Batch 2 Golden Test Dataset	0.470	0.617	0.522
Batch 3 Golden Test Dataset	0.472	0.611	0.512
Batch 4 Golden Test Dataset	0.464	0.75	0.507
Batch 5 Golden Test Dataset	0.444	0.611	0.484

As shown in Tables 9 and 10 above, the highest MRR scores obtained for both TinyBERT and TinyBioBERT models were 0.522 and 0.284, respectively. Although these scores were in the same range as previous BioASQ task 9b competitors, it is recognized that these scores have room for improvement. Since the MRR scores show that the best TinyBERT and TinyBioBERT models had prediction rates of 52% and 28%, future research can be spent improving upon these scores. Other considerations for improving this research include running different TinyBERT and TinyBioBERT models and continuing to tune model hyperparameters.

To understand how these models could be better tuned for future use, log files for both models were reviewed to compare each model's predictions to the exact answers provided by BioASQ's golden datasets. As a result of this analysis three insights were gained: 1) Both models contained predictions that were more specific than the exact answers provided in the golden datasets; 2) Both models contained predictions that did not have special characters that the golden datasets had; and 3) Based on question contexts, there were instances where a question had two correct answers. However, we found that this paper's model predicted the answer that was not in BioASQ's defined exact answer. Table 11 below provides examples of each of these insights. As a result of this analysis, this research concluded that both models were performing better than the model evaluation metrics showed.

Table 11. Examples of Insights Gained from Analyzing Log Files.

Predicted Answer:	Golden Dataset Answer:	Insight Gained:
mineralocorticoid receptor	mineralocorticoid	This paper's models generated predictions that were more specific than the answers provided in the golden dataset.
Tumour necrosis factor (TNF)	TNFα	This paper's models generated predictions that did not have special characters that the golden datasets had.
subtelomeric 9q34.3 deletion	Mutations in the Euchromatic Histone Methyltransferase 1 (EHMT1)	Based on question contexts, there were instances where a question had two correct answers. However, we found that this paper's model predicted the answer that was not in BioASQ's defined exact answer.

5 Discussion

This research successfully built a Question Answering system for healthcare providers using the following distilled BERT models: TinyBERT and TinyBioBERT. Although a functional Question Answering system was built, results for Strict Accuracy showed that this system is not fit for production. To improve upon these results, this research recommends that this system be fitted to other distilled models, and that partnerships be made with organizations that can make EMR data available for research purposes.

Until Strict Accuracy scores improve, it is recommended that healthcare providers use this system cautiously. Specifically, this research recommends that healthcare providers validate the answers they receive from this system to ensure that they are correct. With the help of future research, and more accessible EMR data, this research remains optimistic that Question Answering solutions can be used by healthcare providers to reduce the time they spend searching for answers in patient data.

Although this system is not production-ready, this research showed how Question Answering systems can successfully be applied to closed domain use cases. While Question Answering systems have largely been built to support open domain use cases, this research shows that these systems are just as effective when applied to a specific scope. The Question Answering system in this research applies exclusively to the healthcare industry, however, this system could be adapted to fit another domain such as law.

Challenges encountered in this analysis included understanding the architecture of Transformer models, pre-processing the BioASQ data, and fine-tuning the TinyBERT & TinyBioBERT models. Before any models could be built, understanding the Transformer model architecture was essential. Learning the encoder-decoder structure of these models—coupled with the multi-head attention layers—took careful consideration and time. Understanding the model architecture helped determine how the input data should be formatted and helped identify important hyperparameters that needed to be tuned. After an understanding of Transformer models was gained, the next challenge encountered was pre-processing the BioASQ training and golden datasets. To achieve optimal performance, these datasets had to be modified to match the SQuAD dataset's format. For a detailed discussion on how this was accomplished, please reference sub-section 3.2 of this paper. Finally, the last challenge encountered was fine-tuning the TinyBERT and TinyBioBERT models. The modeling process consisted of modifying training and evaluation functions to produce model results, testing different hyperparameters and hyperparameter values, and testing out different model versions.

6 Ethics

When thinking about the ethics of this Question Answering system, two factors must be considered. The first is model security and privacy. Since this system will be used in a healthcare setting, it is essential that only authorized personnel have access to it. Simple security measures such as role-based access control, and two-factor authentication can be enabled on this system to ensure that best practices for security are met. Regarding model privacy—given that this paper's Question Answering system will contain medical data that is governed by HIPAA—it is critical that patient privacy is enforced.

The second ethical consideration for this research is model bias. Since this Question Answering system was built on a small BioASQ training dataset that is representative of a single population, it should be regularly monitored for bias. This is important because if this system is fitted onto a new population, bias will be introduced.

To prevent model bias, this research recommends fitting this system to an external test set. Using an external test set will help ensure that the system is robust to medical data it was not originally fit on. Should this system show signs of bias, product owners can correct it by supplying the system with a larger training dataset that is more representative of the population it is modeling.

6 Conclusion

In this research, a Question Answering system for healthcare providers was built using the distilled BERT models, TinyBERT and TinyBioBERT. According to this research's evaluation metrics of Strict Accuracy, Lenient Accuracy, and Mean Reciprocal Rank, TinyBERT outperformed TinyBioBERT. However, since both models had low average Strict Accuracy scores, (0.41 & 0.18), it is recommended that additional research be done to improve model performance before this system is productionized.

One limitation of this research was the size of the training dataset. Since it was small, this research hypothesizes that the performance of both the TinyBERT and TinyBioBERT models was negatively impacted. In future research, it is recommended that other researchers fit a larger training dataset to both models and evaluate how model performance is impacted.

Overall, this research demonstrated the effectiveness of distilled BERT models performing Natural Language Processing tasks. Due to their multi-head attention layers, distilled BERT models are uniquely poised to bi-directionally understand word sequences and their contexts. With this capability, this research is optimistic about the insights these models will continue to contribute to the community, especially in the realm of unstructured data.

References

1. Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., & Gupta, M. (2021). COBERT: COVID-19 Question Answering System Using BERT. *Arabian Journal for Science and Engineering (2011),* , 1-11. 10.1007/s13369-021-05810-5

2. Ben Abacha, A., & Demner-Fushman, D. (2016). Recognizing Question Entailment for Medical Question Answering. *AMIA ...Annual Symposium Proceedings; AMIA Annu Symp Proc, 2016*, 310-318.
3. Ben Abacha, A., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC Bioinformatics, 20*(1), 511. 10.1186/s12859-019-3119-4
4. Chaves, A., Kesiku, C., & Begonya Garcia-Zapirain. (2022). Automatic Text Summarization of Biomedical Text Data: A Systematic Review. *Information (Basel), 13*(8), 393. 10.3390/info13080393
5. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
6. Du, Y., Pei, B., Zhao, X., & Ji, J. (2020). Deep scaled dot-product attention based domain adaptation model for biomedical question answering. *Methods (San Diego, Calif.); Methods, 173*, 69-74. 10.1016/j.ymeth.2019.06.024
7. Gao, L., Dai, Z., & Callan, J. (2020). Understanding BERT Rankers Under Distillation. Paper presented at the 10.1145/3409256.3409838
8. Hugging Face.*BERT.* Retrieved 10/17/2022, from https://huggingface.co/docs/transformers/model_doc/bert
9. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2019). TinyBERT: Distilling BERT for Natural Language Understanding.
10. Joshi, S. (2016). How unstructured data will impact precision medicine. *Health Data Management (Online),*
11. Mohammad Mahdinoori. (2022). *Compact Biomedical Transformers.* https://github.com. Retrieved 10/23/2022, from https://github.com/nlpie-research/Compact-Biomedical-Transformers
12. Nentidis, A., Bougiatiotis, K., Krithara, A., & Paliouras, G. (2020). Results of the seventh edition of the BioASQ Challenge. Paper presented at the 10.1007/978-3-030-43887-6_51
13. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Gasco, L., Krallinger, M., & Paliouras, G. (2021). Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. (). Cornell University Library, arXiv.org. 10.1007/978-3-030-85251-1_18
14. Pampari, A., Raghavan, P., Liang, J., & Peng, J. (2018). emrQA: A Large Corpus for Question Answering on Electronic Medical Records.
15. Razzak, M. I., Imran, M., & Xu, G. (2020). Big data analytics for preventive medicine. *Neural Computing & Applications; Neural Comput Appl, 32*(9), 4417-4451. 10.1007/s00521-019-04095-y
16. Rohanian, O., Nouriborji, M., Kouchaki, S., & Clifton, D. A. (2022). On the Effectiveness of Compact Biomedical Transformers.
17. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., & Bhattacharyya, P. (2022). ScienceQA: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries, 23*(3), 289-301. 10.1007/s00799-022-00329-y
18. Sarkar, D. (2016). *Text Analytics with Python A Practical Real-World Approach to Gaining Actionable Insights from your Data* (1st ed.). Apress. 10.1007/978-1-4842-2388-8
19. Sarrouti, M., & Ouatik El Alaoui, S. (2020). SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial Intelligence in Medicine; Artif Intell Med, 102*, 101767. 10.1016/j.artmed.2019.101767