Enhancing Biomedical QA through Task-Adaptive Pretraining: A Transfer Learning Approach

Atharva Pargaonkar
Department of Computer Science
Boise State University
Boise, Idaho, USA
atharvapargaonka@u.boisestate.edu

***Abstract*—The rapid growth of biomedical literature has introduced significant challenges in retrieving accurate and context-aware information for clinical and research applications. This study investigates the effectiveness of task-adaptive pretraining (TAPT) in enhancing the domain robustness of transformer-based language models for binary question answering (QA) in the biomedical field. A RoBERTa model, initially fine-tuned on the general-purpose BoolQ dataset, was adapted to the biomedical domain using TAPT on 6K, 10K, and 15K unlabeled samples from PubMedQA. The adapted models were then fine-tuned on labeled biomedical QA data and evaluated for accuracy and F1 score. Results show that TAPT significantly improves performance over non-adapted baselines, with the best configuration (TAPT-10K) achieving 89.83% accuracy and 0.9408 F1 score. These findings highlight the scalability and utility of TAPT in bridging domain gaps, offering a promising direction for deploying QA systems in specialized domains such as biomedicine.**

***Index Terms*—Task-adaptive pretraining, transfer learning, Question Answering (QA), Natural Language Processing**

I. Introduction

The biomedical domain has experienced an unprecedented surge in the volume of scientific publications, clinical documentation, and research articles. This rapid expansion of textual data presents both an opportunity and a significant challenge. While the abundance of information holds potential for accelerating discoveries, it also contributes to information overload, making it increasingly difficult for clinicians, researchers, and policymakers to access timely and relevant knowledge. Efficiently retrieving contextually accurate information from this vast and specialized corpus is critical for tasks such as clinical decision-making, biomedical research, and the development of healthcare solutions.

Question Answering (QA) systems [1] have emerged as promising tools to address this challenge by delivering direct and concise answers to user queries formulated in natural language. However, the application of QA models in the biomedical field introduces unique complexities. Unlike general-domain texts, biomedical documents are laden with highly technical terminology, frequent abbreviations, and intricate semantic relationships. Models pre-trained on open-domain corpora often fail to capture these domain-specific linguistic patterns, leading to reduced accuracy and reliability in biomedical contexts.

To overcome these limitations, recent advances in Natural Language Processing (NLP) have turned toward transfer learning techniques, particularly the use of transformer-based models such as BERT, RoBERTa, and DistilBERT. These architectures have demonstrated superior performance on QA benchmarks in the general domain by learning deep contextual representations of language.

To address this domain shift, adaptive pretraining techniques such as Domain-Adaptive Pretraining (DAPT) and Task-Adaptive Pretraining (TAPT) have been proposed. DAPT focuses on further training language models on large volumes of unlabeled biomedical text, enabling them to internalize domain-specific knowledge. TAPT refines this approach by continuing the model's pretraining on unlabeled data that closely resembles the format and task structure of the downstream application, such as biomedical question-answer pairs. These strategies enhance the model's understanding of domain-relevant semantics, improving its ability to generalize to specific tasks.

This study explores the effectiveness of TAPT for biomedical question answering by evaluating a RoBERTa model initially fine-tuned on a general-domain QA dataset (BoolQ). The model is further adapted to the biomedical QA domain using TAPT with progressively larger subsets of PubMedQA's unlabeled passages (6K, 10K, and 15K samples). The objective is to assess whether task-specific adaptation can meaningfully improve performance on specialized biomedical questions and how such improvement scales with the amount of adaptation data.

The research is guided by the following questions:

- Can task-adaptive pretraining (TAPT) improve the performance of general-purpose language models on biomedical question answering tasks?
- How does the amount of unlabeled domain-specific data used during TAPT affect downstream QA performance in the biomedical domain?
- How do BERT-based models, including BERT, RoBERTa, and DistilBERT, compare in their ability to generalize from general-domain QA (BoolQ) to biomedical QA (PubMedQA)?

To answer these questions, a series of controlled experiments are conducted involving TAPT, followed by supervised fine-tuning using PubMedQA's labeled dataset. Model performance is evaluated using accuracy and F1 score, while training dynamics such as loss and learning curves are analyzed to understand the impact of different adaptation strategies.

Through this investigation, the study aims to demonstrate the practical utility of TAPT in biomedical QA and offer insights into the optimal use of unlabeled adaptation data for domain transfer. The outcomes of this work have broader implications for improving QA systems deployed in healthcare settings, accelerating literature review processes, and enhancing clinical decision support through reliable, domain-aware language models.

Beyond research implications, the ability to adapt question answering models to the biomedical domain has tangible real-world utility. In fast-paced clinical environments, such systems can serve as intelligent assistants for first responders or emergency care teams by providing quick, evidence-based answers to medically critical questions. When integrated into mobile or web applications, these models could follow structured decision-making flows—akin to if-else logic trees—to support preliminary diagnoses, treatment triage, or patient risk assessment. Such tools could significantly improve response time, reduce cognitive load on medical staff, and facilitate better decision-making under pressure in high-stakes healthcare settings

The successful execution of this project was greatly supported by the foundational knowledge gained through prior coursework in the Computer Science Master's program, including Natural Language Processing, Machine Learning, Data Science, and Large-Scale Data Analytics. These courses provided a strong theoretical grounding in modern AI techniques and the ability to process large-scale datasets—skills that were critical for designing and implementing transformer-based language models and domain adaptation strategies. The project directly reflects the first Program Learning Outcome (PLO), which emphasizes the application of computer science theory to define and solve complex problems. From selecting appropriate models and formulating transfer learning pipelines to interpreting experimental results, every phase of this study involved analytically designing and executing technically sound solutions.

The second PLO—centered on communicating technical work and its scientific and societal impact—was met through the structured documentation of methodology, analysis of results using visualizations and performance metrics, and reflection on how such QA systems could enhance clinical decision-making and emergency response in healthcare settings. Courses like Advanced Software Engineering further developed my ability to present complex information in clear, research-oriented writing. Lastly, the third PLO, which involves engaging in self-directed learning, was continuously demonstrated throughout the project. Exploring cutting-edge techniques, experimenting with different model architectures, and managing domain-specific challenges required independent research, experimentation, and adaptation beyond standard coursework.

II. Related Work

A. Foundations of Transfer Learning and Domain Adaptation

Transfer Learning (TL) refers to techniques that leverage knowledge from a source domain (or task) to improve learning in a target domain where training data is limited [2]. A core assumption in traditional machine learning is that training and test data are drawn from the same distribution [3]. In reality, differences in vocabulary, style, or content between domains lead to domain shift (also called dataset shift), which causes models trained on one data distribution to degrade in performance when applied to another [3]. Domain adaptation (DA) is a special case of transfer learning that tackles this scenario: it assumes the feature space and label space are the same in source and target, and only the data distributions differ [4].

Homogeneous vs. Heterogeneous Transfer Learning: In homogeneous transfer learning, the source and target domains share the same feature space and label space (X_s = X_t and Y_s = Y_t) [2]. The primary challenge here is to overcome distribution disparity (domain shift) while using the same representations. Most classic domain adaptation research (a form of homogeneous TL) focuses on mitigating the drop in accuracy caused by differing data distributions across domains. In contrast, heterogeneous transfer learning deals with different feature and/or label spaces between source and target. Heterogeneous settings are more complex, often requiring learning mappings or shared latent spaces to bridge the gap between domains with completely different data representations (e.g. transferring between text and image domains, or between different languages).

B. Domain-Adaptive and Task-Adaptive Pretraining (DAPT & TAPT)

Pretrained language models like BERT and RoBERTa, trained on massive general-domain corpora, provide a strong starting point for NLP tasks. Domain-Adaptive Pretraining (DAPT) refers to an additional phase of unsupervised language-model pretraining on text from a specific target domain before fine-tuning on a task [5]. Gururangan et al. (2020) showed that in-domain pretraining consistently yields performance gains on downstream tasks in the target domain, even when the base model was a large one like RoBERTa. Notably, these gains were observed in both high-resource settings (plenty of task data) and low-resource settings, underscoring that domain mismatch is an important factor to address regardless of task data size.

C. Domain-Specific Pretrained Models in Biomedical NLP

Following the success of BERT [6], the biomedical NLP community identified that general-purpose language models could be significantly improved through domain-adaptive pretraining (DAPT) or full pretraining on biomedical corpora. This insight led to the development of numerous domain-specific transformer-based models, each tailored to the unique characteristics of scientific and clinical texts.

BioBERT [7] is among the first such models. Built upon BERT-base, it was further pretrained on PubMed abstracts and PMC full-text articles. While it retains BERT's original vocabulary and architecture, the exposure to biomedical terminology during DAPT allows BioBERT to develop specialized representations of domain-specific entities such as gene names and medical terms. Upon release, BioBERT achieved state-of-the-art results on tasks including named entity recognition (NER), relation extraction, and question answering (QA), outperforming general-domain BERT in biomedical QA benchmarks like BioASQ and PubMedQA.

SciBERT [8] was trained from scratch on 1.14 million scientific papers from Semantic Scholar. Approximately 18% of the corpus consists of biomedical literature, while the remainder covers broader scientific disciplines such as computer science. Unlike BioBERT, SciBERT introduces a domain-specific vocabulary optimized for scientific terms. While it is not exclusively biomedical, SciBERT demonstrates robust performance on a wide range of science and biomedical NLP tasks, proving the efficacy of mixed-domain pretraining when combined with an aligned tokenizer.

PubMedBERT [9] takes domain adaptation further by training a BERT model from scratch solely on PubMed abstracts. It abandons BERT's original vocabulary and instead learns a new one tailored for biomedical texts. This full-domain specialization results in superior performance across many biomedical NLP tasks. For instance, on the BLURB benchmark—covering NER, sentence similarity, document classification, and QA—PubMedBERT surpasses all previously established baselines, demonstrating the impact of vocabulary and corpus alignment.

Other notable contributions include **ClinicalBERT** [10] and **BlueBERT** [11], both of which are adapted for clinical texts such as MIMIC-III electronic health records. These models were either initialized from BERT or BioBERT and further trained on clinical notes. While they show performance gains on clinical tasks compared to general BERT, they typically lag behind models like PubMedBERT that benefit from cleaner and larger biomedical training corpora.

D. Transfer Learning for Biomedical Question Answering

Biomedical question answering (QA) presents a complex challenge, particularly due to the scarcity of annotated training data compared to the open domain. The domain involves specialized terminology, dense scientific language, and the necessity for nuanced reasoning over research articles. As noted by Athenikos and Han [12], biomedical QA often deals with vast corpora yet limited supervision, making it an ideal candidate for transfer learning techniques.

Two major types of biomedical QA tasks are commonly studied: factoid QA, as featured in the BioASQ challenge, and yes/no QA, exemplified by datasets like PubMedQA. Early attempts to adapt open-domain QA models to the biomedical domain include the work by Wiese et al. [13], where a neural QA model trained on the large-scale SQuAD dataset was fine-tuned for BioASQ. Despite BioASQ's limited training set (fewer than 900 annotated factoid questions), the authors achieved state-of-the-art performance by leveraging biomedical word embeddings and fine-tuning strategies, without relying on ontologies or handcrafted features.

PubMedQA introduced in 2019 [14], targets yes/no/maybe QA in the biomedical field. Each instance consists of a research question derived from a PubMed article title, an abstract as the context, and the conclusion sentence as a long answer. The task is to infer a short answer—"yes," "no," or "maybe"—from the evidence presented in the abstract. The dataset provides 1,000 expert-labeled QA pairs, 61,000 unlabeled instances, and 211,000 artificially generated questions for semi-supervised learning. It serves as the biomedical counterpart to the general-domain BoolQ dataset.

Transfer learning played a critical role in the initial benchmarks for PubMedQA. The authors reported that BioBERT, when fine-tuned through a multi-phase training strategy, achieved 68.1% classification accuracy—outperforming the majority-class baseline by over 13 percentage points and approaching single-human performance (78%) [14]. This involved first adapting the model on synthetic and unlabeled data (self-supervised and weakly supervised stages), followed by fine-tuning on the 1k expert-labeled set.

III. Dataset Description

This study utilizes two publicly available datasets for binary question answering: **BoolQ** and **PubMedQA**. These datasets represent distinct domains—general and biomedical—and serve complementary purposes in this work. BoolQ acts as the source domain for transfer learning, while PubMedQA functions as the target domain where domain adaptation techniques are evaluated.

A. BoolQ: General-Domain QA Dataset

BoolQ (Boolean Questions) is a question-answering dataset introduced by Clark et al. (2019) [15], designed to evaluate a model's ability to answer yes/no questions given a supporting passage. It is constructed from naturally occurring questions collected from Google search queries and passages extracted from Wikipedia articles.

- **Structure:** Each instance in BoolQ is a triplet containing:
  - `question`: a naturally phrased yes/no question
  - `passage`: a paragraph from a Wikipedia article
  - `answer`: a boolean value (true or false)
- **Dataset Size:**
  - Training set: 9,427 samples
  - Development set: 3,270 samples
  - Test set: 3,245 samples (unlabeled)
- **Sample Instance:**

Question: *Does ethanol take more energy make that produces?*

Passage: *All biomass goes through at least some of these steps: it needs to be grown, collected, dried, fermented, distilled...*

Answer: `False`

- **Preprocessing:**
  - Removed extra whitespaces and ensured consistent formatting.
  - Retained only relevant columns (`question`, `passage`, `answer`).
  - Applied truncation and padding to maintain input consistency.
- **Role in Pipeline:** BoolQ is used to fine-tune pre-trained models (e.g., RoBERTa) in a general QA setting. This simulates a real-world scenario where models are trained on a large general-domain QA dataset and later adapted to a new domain (biomedical). The fine-tuned model serves as the starting point for domain adaptation using TAPT on PubMedQA.

B. PubMedQA: Biomedical QA Dataset

PubMedQA is a benchmark dataset specifically designed for biomedical yes/no/maybe question answering. It contains questions derived from real-world biomedical research articles indexed in PubMed, making it highly suitable for evaluating domain-specific language models.

- **Variants Used:**
  - `ori_pqal.json`: 1,000 human-annotated QA pairs (used as labeled data)
  - `ori_pqaa.json`: 211,269 automatically generated QA pairs with noisy answers (used for TAPT)
  - `ori_pqau.json`: 61,249 unlabeled QA instances (used for TAPT)
- **Structure:** Each instance contains:
  - `pubmed_id`: unique article identifier
  - `QUESTION`: a biomedical yes/no/maybe question
  - `CONTEXTS`: a list of sentences or abstract from the article
  - `final_decision`: the annotated answer
- **Sample Instance:**

pubmed_id: 14499029

Question: *Is laparoscopic radical prostatectomy better than traditional retropubic radical prostatectomy?*

Passage: *The clinical and pathological data obtained in 50 consecutive patients who underwent retropubic radical prostatectomy (RRP) from January 2001 to December 2001 ...*
Answer: Yes

- **Preprocessing:**

1) *Context Processing:* The `CONTEXTS` list is concatenated into a single `passage` string to match BoolQ format.
2) *Answer Mapping:* The `final_decision` field is mapped to binary: `yes` → 1 (True), `no` or `maybe` → 0 (False). This binary simplification ensures compatibility with BoolQ-style Boolean QA tasks.
3) *Data Integration:* The labeled `pqal` set (1,000 samples) is combined with a random 5,000-sample subset from `pqaa`, forming a combined labeled dataset of 6,000 samples. This set is split into 80% training and 20% validation.

- **Role in Pipeline:**
  - *Unlabeled subsets* (`pqaa`, `pqau`) are used for **Task-Adaptive Pretraining (TAPT)**.
  - *Labeled subsets* (`pqal` + sampled `pqaa`) are used for **supervised fine-tuning and evaluation**.

By using this two-dataset setup, the project evaluates how well general-purpose language models can be adapted to domain-specific tasks using TAPT and fine-tuning, and how performance varies with the amount of domain-specific unlabeled data. The consistent yes/no format in both datasets enables a controlled and comparable experimental framework across domains.

IV. Model Background

This study leverages transformer-based pre-trained language models to address the task of binary question answering in the biomedical domain. Among the available options, RoBERTa (Robustly Optimized BERT Pretraining Approach) was selected as the primary model based on its strong performance in initial experiments conducted on the general-domain BoolQ dataset. To establish a well-rounded understanding of the modeling approach, this section outlines the underlying architecture of these models, their pretraining strategies, and the rationale behind the choice of RoBERTa as the principal model.

A. Transformer Architecture

The transformer architecture, introduced by Vaswani et al. (2017), forms the foundation of modern language models. It replaces traditional sequential processing mechanisms found in RNNs and LSTMs with a parallelizable mechanism known as self-attention. This mechanism allows the model to weigh the importance of different words in a sentence relative to one another, regardless of their positions. As a result, transformers can effectively model long-range dependencies and contextual relationships within text, which are essential for nuanced language understanding.

Transformers consist of stacked layers, each comprising multi-head self-attention and position-wise feedforward networks, augmented by residual connections and layer normalization. These architectural innovations enable deep models to be trained efficiently and achieve superior performance on various NLP tasks.

B. Pre-trained Language Models

Pre-trained language models are trained on large-scale unlabeled corpora using objectives such as Masked Language Modeling (MLM), where a portion of the input tokens are masked and the model learns to predict them. This process allows the model to acquire an understanding of grammar, syntax, and semantics without explicit supervision. After pretraining, these models are fine-tuned on downstream tasks using labeled data, allowing them to specialize in specific applications such as question answering.

C. BERT: Bidirectional Encoder Representations from Transformers

BERT, introduced by Devlin et al. (2018), marked a paradigm shift in NLP by enabling deep bidirectional context understanding. It uses two pretraining tasks: Masked Language Modeling and Next Sentence Prediction (NSP). MLM trains the model to predict masked tokens using both left and right context, while NSP helps it understand sentence-level relationships. BERT has demonstrated strong performance across a range of NLP benchmarks.

In this study, BERT was used as a baseline model to assess the effectiveness of domain adaptation techniques. Despite its foundational importance, BERT's performance on downstream tasks in specialized domains like biomedicine tends to be limited without further adaptation.

D. DistilBERT: Lightweight Transformer Model

DistilBERT, developed by Sanh et al. (2019) [16], is a distilled version of BERT designed to reduce model size and inference time while retaining a significant portion of BERT's performance. It is trained using a technique called knowledge distillation, where a smaller model (student) learns to replicate the behavior of a larger pre-trained model (teacher). DistilBERT maintains approximately 97% of BERT's performance while being 40% smaller and 60% faster.

In this work, DistilBERT was evaluated to compare its performance against larger models. Its role was primarily to serve as a lightweight baseline and assess the trade-offs between model complexity and predictive accuracy in binary QA tasks.

E. RoBERTa: Robustly Optimized BERT Approach

RoBERTa, proposed by Liu et al. (2019) [17], is an optimized variant of BERT that removes the NSP objective and introduces several key improvements: larger mini-batches, longer training time, dynamic masking, and training on a much larger corpus. These modifications lead to significant performance gains on various NLP tasks.

Preliminary experiments on the BoolQ dataset revealed that RoBERTa outperformed both BERT and DistilBERT, especially in terms of F1 score. This motivated its selection as the primary model for the subsequent domain adaptation and fine-tuning experiments conducted on the PubMedQA dataset.

V. Study Design and Methodology

The goal of this study is to investigate how task-adaptive pretraining (TAPT) can enhance the performance of transformer-based language models for binary question answering (QA) in the biomedical domain. Specifically, the work explores the transfer of knowledge from a general-domain QA dataset (BoolQ) to a biomedical-domain QA dataset (PubMedQA) using various stages of pretraining and fine-tuning. To systematically address the aforementioned research questions, the study is divided into several phases, as described below:

A. Phase 1: Dataset Preparation and Preprocessing

The study utilizes two main datasets—BoolQ as the source domain and PubMedQA as the target domain. BoolQ contains general yes/no questions with associated passages, while PubMedQA features biomedical yes/no/maybe questions with abstracts from PubMed articles. The choice of BoolQ is motivated by its relevance to binary QA and its compatibility with the format of PubMedQA.

- **BoolQ:** Preprocessing involved loading and converting JSONL files to CSV, followed by tokenization using standard HuggingFace tokenizers. The dataset was divided into training, development, and test sets, and the input format was adjusted to match that of PubMedQA to ensure compatibility in cross-domain evaluation.
- **PubMedQA:** Three variants : `ori_pqal.json` (labeled), `ori_pqaa.json` (artificially generated), and `ori_pqau.json` were parsed to construct training and TAPT corpora. Preprocessing included flattening multi-sentence abstracts into passages, mapping multi-class answers to binary labels, and splitting the labeled data into train and dev sets.

The consistency in input format between the two datasets ensured that pretrained models could be evaluated on PubMedQA without architectural modifications. This step is crucial as it ensures the model receives high-quality input data aligned with the task's requirements and enables domain transfer via a cleanly structured learning setup.

B. Phase 2: General-Domain QA Fine-Tuning

To establish a strong baseline and simulate a realistic transfer scenario, three transformer-based models—BERT, DistilBERT, and RoBERTa—were initially fine-tuned on BoolQ. This step mirrors a practical situation where a model trained on a large, general-domain corpus is repurposed for a specialized domain.

- **Motivation:** Fine-tuning on BoolQ helps the models internalize the structure of binary QA tasks, making them capable of handling yes/no questions.
- **Outcome:** RoBERTa significantly outperformed BERT and DistilBERT in terms of F1 score on the BoolQ dev set, justifying its selection as the primary model for further domain adaptation.

C. Phase 3: Task-Adaptive Pretraining (TAPT)

Following general-domain fine-tuning, RoBERTa was subjected to TAPT on the PubMedQA dataset using varying amounts of unlabeled biomedical text (6K, 10K, and 15K samples). TAPT involves continued masked language modeling (MLM), a self-supervised learning objective where random tokens in a passage are masked, and the model is trained to predict them using the surrounding context. This approach allows the model to refine its understanding of domain-specific vocabulary, sentence structures, and contextual patterns, even without labeled data.

In this study, MLM was conducted using the HuggingFace Trainer API with custom data collators that dynamically masked tokens during training. Each passage was tokenized and formatted into input sequences with padding and attention masks, and the model was trained for three epochs for each dataset size. This step was essential to help the general-domain model adapt to the nuanced language used in biomedical literature, thereby improving downstream question answering performance on PubMedQA.

D. Phase 4: Supervised Fine-Tuning on PubMedQA

After TAPT, the RoBERTa models were fine-tuned on the labeled portion of PubMedQA, consisting of 1,000 training and 250 development samples. This supervised phase employed a binary classification head to predict whether the answer to a given biomedical question was ’yes’ or ’no’ based on the associated passage.

The fine-tuning process used cross-entropy loss as the optimization objective, which measures the difference between the predicted and actual class distributions. Training was carried out using the AdamW optimizer—a variant of Adam that decouples weight decay for better generalization—alongside a linear learning rate scheduler. Additional techniques such as gradient clipping were used to stabilize training and prevent exploding gradients.

- **Motivation:** Supervised fine-tuning teaches the model to align its learned representations from TAPT with the actual downstream QA task, providing essential task-specific supervision.
- **Training Dynamics:** Models were fine-tuned for three epochs, and dev set performance was tracked using accuracy and F1 score. The best checkpoint was retained based on maximum F1.
- **Comparison Strategy:** All three TAPT variants (6K, 10K, 15K) were fine-tuned under identical settings to enable fair and interpretable performance comparisons.
- **Insight:** This step reveals how well domain familiarity acquired via TAPT translates into task-specific competence in biomedical QA.

E. Final Evaluation and Cross-Model Comparison

After supervised training, all models—including BoolQ-only, TAPT-6K, TAPT-10K, and TAPT-15K—were evaluated on the PubMedQA development set. Performance metrics included accuracy and F1 score.

The evaluation also involved generating visualizations such as training loss curves, epoch-wise performance trends, and summary charts to illustrate the comparative effects of TAPT at different data scales. These analyses provided empirical evidence for how increasing TAPT data size improves task performance.

VI. Results and Analysis

A. Overview of Evaluation Strategy

The performance of all models was evaluated on the PubMedQA dev set using two key metrics: ***Accuracy and F1 Score***. These metrics were selected to reflect both the correctness and balance of predictions in a binary classification setting. The experiments were designed to compare:

- **Baseline general-domain models** (BERT, DistilBERT, RoBERTa fine-tuned on BoolQ)
- **RoBERTa without domain adaptation** (directly evaluated on PubMedQA)
- **RoBERTa with TAPT** using 6K, 10K, and 15K unlabeled PubMedQA samples
- **Post-TAPT fine-tuning on PubMedQA labeled data** to assess domain adaptation effectiveness

Each variant underwent the same fine-tuning and evaluation pipeline to ensure consistency.

B. Model Performance Comparison

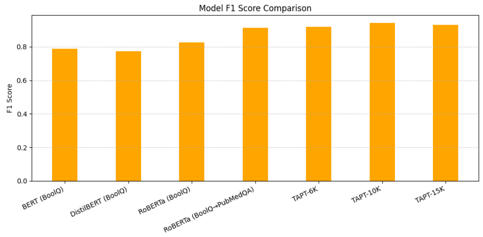

Fig. 1. F1 score comparison across all models and adaptation strategies.

RoBERTa models adapted through TAPT significantly outperformed the general-domain baselines in terms of F1 score. Notably, the TAPT-10K model achieved the highest F1 score of 0.9408, as illustrated in Figure 1. Accuracy trends closely followed F1 scores, with TAPT-10K and TAPT-15K delivering the best results. The complete accuracy comparison across all models is shown in Figure 2.

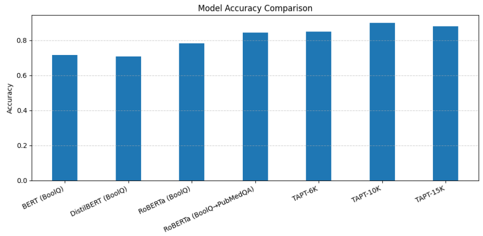

Fig. 2. Accuracy comparison across all models and adaptation strategies.

A comparative summary of all model performances is shown in Table II. It demonstrates the progressive improvement as TAPT and fine-tuning techniques were incorporated.

TABLE I
PERFORMANCE COMPARISON OF BASELINE MODELS ON BOOLQ

Model	Accuracy	F1 Score
BERT (BoolQ)	0.7153	0.7893
DistilBERT (BoolQ)	0.7073	0.7735
RoBERTa (BoolQ)	**0.7835**	**0.8261**

TABLE II
PERFORMANCE COMPARISON ON PUBMEDQA DEV SET

Model	Accuracy	F1 Score
BERT (BoolQ)	0.7153	0.7893
DistilBERT (BoolQ)	0.7073	0.7735
RoBERTa (BoolQ)	0.7835	0.8261
RoBERTa (→ PubMedQA)	0.8442	0.9131
TAPT-6K + Fine-Tuning	0.8483	0.9179
TAPT-10K + Fine-Tuning	**0.8983**	**0.9408**
TAPT-15K + Fine-Tuning	0.8792	0.9296

C. General-Domain Baseline Models (BoolQ Only)

The baseline models trained exclusively on the general-domain BoolQ dataset show a clear performance gap when evaluated on biomedical questions from PubMedQA:

- **BERT (BoolQ)** achieved an accuracy of 71.53% and F1 score of 78.93%, while **DistilBERT** lagged slightly behind (70.73% accuracy, 77.35% F1).
- **RoBERTa (BoolQ)** performed better among the three with 78.35% accuracy and 82.61% F1.

These results suggest that while transformer-based models can transfer some general QA capabilities, they fall short in the biomedical domain due to unfamiliar terminology and contextual gaps—validating the need for domain adaptation.

Insight (RQ1): General QA models exhibit limited effectiveness on specialized biomedical questions, emphasizing the necessity for task-adaptive techniques.

D. TAPT and Fine-Tuning Improvements

To bridge the domain gap, we applied task-adaptive pre-training (TAPT) on unlabeled PubMedQA abstracts using RoBERTa, followed by supervised fine-tuning on labeled biomedical data. Below is a summary of performance improvements:

- **TAPT-6K** improved marginally over RoBERTa baseline (Accuracy: 84.83%, F1: 91.79%).
- **TAPT-10K** achieved the best performance overall (Accuracy: 89.83%, F1: 94.08%), suggesting a sweet spot in terms of domain data scale.
- **TAPT-15K**, while still strong (Accuracy: 87.92%, F1: 92.96%), did not surpass 10K, indicating diminishing returns at higher TAPT scales.

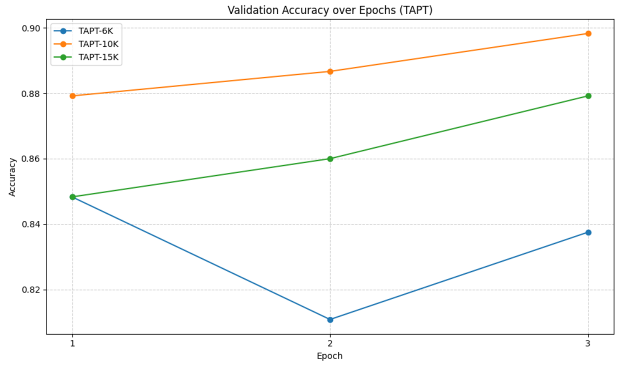

Fig. 3. Validation accuracy over epochs for TAPT-6K, 10K, and 15K.

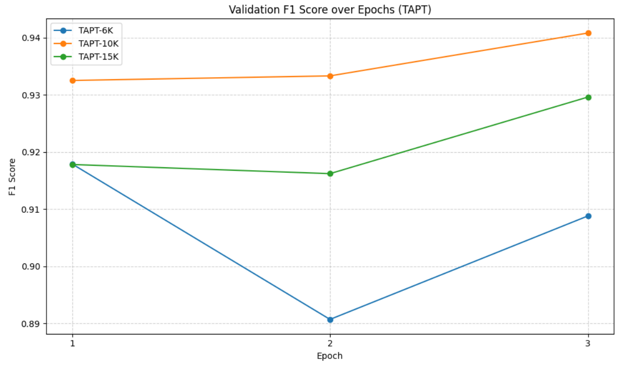

Fig. 4. Validation F1 score over epochs for TAPT-6K, 10K, and 15K.

Validation accuracy curves (Figure 3) demonstrate consistent improvements across epochs, particularly for TAPT-10K and TAPT-15K, suggesting better generalization to biomedical QA. As shown in Figure 4, the F1 scores increased steadily with training, with the highest values recorded in the final epoch of TAPT-10K and TAPT-15K fine-tuning.

Insight (RQ2): Performance gains plateau after a certain TAPT corpus size. While more data generally helps, TAPT-10K yielded the most balanced results. Further increases may introduce noise or redundancy.

A heatmap of model performance provides a visual snapshot of accuracy and F1 scores across configurations, highlighting the improvements gained through domain adaptation (see Figure 5).

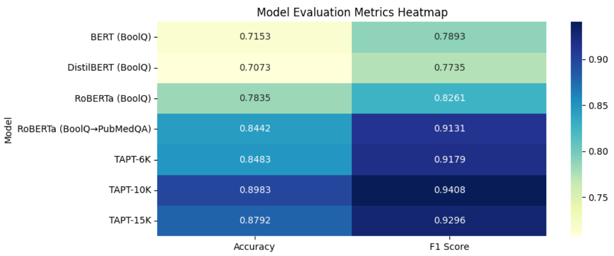

Fig. 5. Heatmap showing accuracy and F1 performance of models across stages.

E. Training Loss and Learning Dynamics

In addition to evaluation metrics like accuracy and F1 score, step-wise training loss values were recorded during task-adaptive pretraining (TAPT) across all three data scales—6K, 10K, and 15K. These losses, computed over masked language modeling (MLM), provide insight into how well the models learned domain-specific representations over time.

- **TAPT-6K** loss decreased from an initial value of **6.38** at step 100 to **1.57** by step 2200. The loss reduction was steady, but exhibited minor fluctuations between steps 1700–2000, suggesting slower convergence and possibly a limited variety of examples in the smaller corpus.
- **TAPT-10K** showed a very stable and consistent decline, starting from **6.46** at step 100 and reaching **1.48** by step 3700. The curve had smoother transitions with less jitter, indicating that the additional data contributed to better generalization and less overfitting. The stable drop across such a large number of steps reflects that RoBERTa was able to extract increasingly complex biomedical features as the training progressed.
- **TAPT-15K** began at a significantly lower loss (**1.42**) due to continuation from a previously fine-tuned checkpoint, and further declined to around **1.13** by step 5600. The curve remained shallow with occasional plateaus and even slight increases (e.g., steps 2200–2500), possibly due to data redundancy or noise in the larger unlabeled set. This might explain why performance gains plateaued despite a longer and larger-scale pretraining.

Step-wise training loss curves reveal that TAPT-10K and TAPT-15K models consistently converged faster and more smoothly compared to TAPT-6K, indicating more effective domain adaptation (refer to Figure 6).

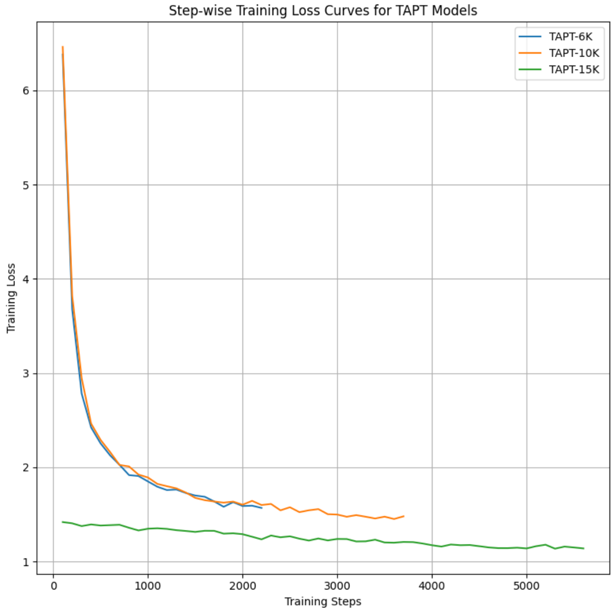

Fig. 6. Step-wise training loss during TAPT for 6K, 10K, and 15K samples.

These training dynamics highlight a few important trends:

- **TAPT-10K** provided the most effective trade-off between loss minimization and generalization, as evidenced by its smooth curve and highest final F1 score (94.08).
- **TAPT-6K** was sufficient to introduce domain familiarity but lacked data diversity to yield optimal results.
- **TAPT-15K**, despite having the longest training and lowest final loss, did not produce the best evaluation scores, reinforcing that more pretraining data does not always lead to better task performance.

Insight (RQ3) : This result directly supports RQ3, as TAPT models significantly outperformed both the general-domain (BoolQ-only) and direct fine-tuning baselines, illustrating the superior impact of task-adaptive pretraining in specialized biomedical QA.

VII. Conclusion and Future Work

This study examined the effectiveness of transfer learning and task-adaptive pretraining (TAPT) in enhancing biomedical question answering performance using transformer-based models. Faced with the challenge of limited annotated biomedical data, the research adopted a multi-phase approach involving initial fine-tuning on a general-domain QA dataset (BoolQ), followed by TAPT using varying amounts of unlabeled PubMedQA data, and culminating in supervised fine-tuning on a small labeled biomedical corpus.

The results clearly demonstrate that models pre-adapted to the biomedical domain via TAPT significantly outperform both general-domain baselines and models directly fine-tuned on the biomedical data. The RoBERTa model, after undergoing TAPT with 10K samples, achieved the highest performance on the PubMedQA development set with an F1 score of 0.9408 and an accuracy of 0.8983, compared to the non-adapted RoBERTa baseline which achieved an F1 score of 0.9131. Despite the strong results, this work is not without limitations. The binary framing of the problem excluded "maybe" responses from PubMedQA. Moreover, the TAPT process is computationally demanding and may not scale efficiently for larger language models or full pretraining scenarios.

Future research could address these aspects by incorporating the full ternary label structure of PubMedQA and experimenting with generative models capable of handling open-ended or nuanced answers. Additionally, leveraging semi-supervised learning strategies to utilize the vast amount of unlabeled biomedical text more efficiently could further enhance performance. Expanding error analysis and interpretability efforts may also help pinpoint the remaining challenges in biomedical QA and inform future model designs.

From an application standpoint, integrating such QA systems into user-friendly web platforms or mobile applications could support clinicians, researchers, and even patients in accessing fast, evidence-backed answers to biomedical queries. A future direction could involve developing intelligent assistant tools for clinical diagnosis that interactively respond to user queries in real time using domain-adapted models. Such tools have the potential to access biomedical knowledge, assist with early diagnosis, and ultimately improve healthcare outcomes.

REFERENCES

[1] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, 2012.

[2] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *Journal of Big Data*, vol. 4, pp. 1–42, 2017.

[3] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in nlp—a survey," *arXiv preprint arXiv:2006.00632*, 2020.

[4] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," *arXiv preprint arXiv:1812.11806*, 2018.

[5] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[8] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[9] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[10] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[11] Z. Peng, W. Zhang, N. Han, X. Fang, P. Kang, and L. Teng, "Active transfer learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1022–1036, 2019.

[12] S. J. Athenikos and H. Han, "Biomedical question answering: A survey," *Computer methods and programs in biomedicine*, vol. 99, no. 1, pp. 1–24, 2010.

[13] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," *arXiv preprint arXiv:1706.03610*, 2017.

[14] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[15] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *arXiv preprint arXiv:1905.10044*, 2019.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.