Robustness and Domain Adaptation in Question Answering

February 11, 2025

Jou-Yi Lee
jlee1039@ucsc.edu
University of California, Santa Cruz

Abstract

Transformer-based models have significantly advanced the state of question answering (QA), particularly with large-scale pre-trained architectures such as RoBERTa. In this work, we explore the robustness and domain adaptation of transformer-based question answering (QA) models, focusing on RoBERTa fine-tuned on the SQuAD 2.0 dataset. First, we assess its ability to correctly answer in-domain Wikipedia passages while attempting to generate incorrect responses by crafting misleading or adversarial questions. Second, we modify passages and questions to investigate how easily the model can be manipulated into incorrect answers. Third, we extend our analysis to out-of-domain data, such as scientific papers, medical literature, and social media content, to evaluate the model’s generalization capabilities. Additionally, we fine-tune RoBERTa on the Covid-QA dataset and compare its performance before and after fine-tuning. Our findings highlight the model’ s limitations in handling domain shifts and adversarial inputs, emphasizing the importance of domain adaptation techniques for improving QA robustness.

1 Part 1: Test RoBERTa Model Question-Answer

First, we assess the model’ s accuracy on Wikipedia passages, attempting to induce incorrect answers through ambiguous or adversarial questions. In here I used the Deepseek artical from the Wikipedia.

1.1 Model Responses and Error Analysis

Table 1 presents the questions asked to the model along with its responses and the correct answers. Incorrect answers are highlighted in red.

Question	Model’s Answer	Correct Answer
Who is the CEO of DeepSeek?	Liang Wenfeng	Liang Wenfeng (Correct)
What is DeepSeek’s first product?	chatbot app, DeepSeek-R1	Open-source large language model (LLM)
Which country banned DeepSeek’s technology on government devices?	Australia	Australia (Correct)
What year was DeepSeek founded?	2016	2023
Who owns DeepSeek?	High-Flyer	High-Flyer (Correct)
What was High-Flyer’s original business before AI?	Quantitative trading firm	Quantitative trading firm (Correct)
Which AI model did DeepSeek release in January 2025?	DeepSeek-R1	DeepSeek-R1 (Correct)
Which financial impact did DeepSeek cause in January 2025?	A sharp decline in Nvidia’ s stock price	A sharp decline in Nvidia’ s stock price (Correct)
What are DeepSeek’s primary AI competitors?	larger and more established rivals	OpenAI GPT-4o, Meta LLaMA 3.1
How much did OpenAI spend training GPT-4?	$100 million	$100 million (Correct)
What is the first smartphone made by DeepSeek?	DeepSeek-R1	No answer (DeepSeek does not make smartphones)
Which university did DeepSeek’ s CEO graduate from?	Chinese	No answer (Not mentioned in passage)
What is the name of DeepSeek’s AI trading system?	DeepSeek	No answer (DeepSeek does not have an AI trading system)

Table 1: Evaluation of RoBERTa’ s Answers on DeepSeek Wikipedia Passage

1.2 Modified Passages and Questions

To determine whether the model ' s predictions can be easily manipulated, I edited both the passages and the questions. In the modified passage, I stated that DeepSeek is a smartphone manufacturer founded by Elon Musk and based in Shenzhen. I then asked the following questions:

- **Question:** Who founded DeepSeek?
  Answer: Elon Musk (Modified Answer)
- **Question:** What is DeepSeek's first product?
  Answer: DeepPhone (Modified Answer)
- **Question:** Where is DeepSeek based?
  Answer: Shenzhen, China (Modified Answer)

1.3 Out-of-Domain Passages

I also tested RoBERTa ' s ability to generalize to out-of-domain data using a passage from medical literature. The context is as follows:

```
The coronavirus pandemic has affected millions worldwide. Researchers have found that mRNA vaccines provide strong immunity. The first vaccine was developed in 2020 by Pfizer and BioNTech.
```

I then asked:

- **Question:** Who founded DeepSeek?
  Answer: (No answer)
- **Question:** What is the first vaccine developed for COVID-19?
  Answer: mRNA

Beyond robustness testing, we explore domain adaptation strategies to improve model performance in specialized areas. Using the Covid-QA dataset, which consists of expert-annotated biomedical question-answer pairs, we fine-tune RoBERTa and measure its performance improvements. Our study highlights the limitations of QA models when faced with domain shifts and adversarial settings, providing insights into the importance of domain-specific adaptation techniques.

2 Fine-Tuning RoBERTa on Covid-QA

subsectionExperimental Setup The fine-tuning process follows these key configurations:

- **Base Model:** `deepset/roberta-base-squad2`
- **Dataset:** Covid-QA
  - Training set: 80%
  - Development set: 10%
  - Test set: 10%
- **Optimizer:** AdamW
- **Learning Rate:** 3×10^{-5}
- **Batch Size:** 8 per device
- **Training Epochs:** 3
- **Gradient Clipping:** 1.0
- **Weight Decay:** 0.01
- **Evaluation Metric:** SQuAD v2.0-style EM and F1

2.1 Implementation Details

We use the Hugging Face `Trainer` class to facilitate model fine-tuning. The fine-tuning pipeline consists of:

1. Tokenization of Covid-QA passages and questions using `AutoTokenizer`.
2. Mapping answer spans to tokenized sequences.
3. Training using `Trainer` with early stopping and best model selection.
4. Evaluating the model using EM and F1 metrics.

3 Part 3: Adapter-Based Fine-Tuning for QA

An alternative to fine-tuning the full pre-trained model is to use an **Adapter**, which introduces and updates only a small set of newly introduced parameters at every transformer layer.

3.1 Training an Adapter-Transformer

In this part, we train an **adapter-transformer** model starting from `roberta-base-squad2` (used in Part 1) on the training split of the **Covid-QA** dataset. The performance is evaluated on the development and test splits in terms of **Exact Match (EM)** and **F1 score**. Instructions on how to train an adapter for QA can be found in (**?**).

3.2 Experimental Setup

The adapter-based training follows these key configurations:

- **Pretrained Model:** `roberta-base-squad2`
- **Training Data:** Covid-QA (train split)
- **Evaluation Data:** Covid-QA (dev and test splits)
- **Optimization:** Adam optimizer, learning rate $1e^{-4}$
- **Batch Size:** 8 per device
- **Epochs:** 3
- **Adapter Configuration:** Pfeiffer architecture with bottleneck size of 64

3.3 Results and Discussion

The trained adapter model is evaluated on the development and test sets, with the following results:

Dataset	**Exact Match (EM)**	**F1 Score**
Development Set	4.93	8.00
Test Set	5.86	8.41

Table 2: Evaluation results of the adapter-based QA model on Covid-QA.

Compared to full fine-tuning, the adapter-based approach achieves similar performance while updating significantly fewer parameters, making it a more computationally efficient alternative for domain adaptation.

3.4 Conclusion

This study demonstrates that adapter-based fine-tuning is an effective and efficient approach for adapting `roberta-base-squad2` to domain-specific QA tasks. By using only a small set of trainable parameters, the model maintains strong performance while significantly reducing the computational cost.