NLP 203: Assignment 2

Ishika Kulkarni
ikulkar1@ucsc.edu

Abstract

This report addresses the robustness and domain adaptation of the RoBERTa model for question-answering tasks. The first problem we work on is testing robustness by evaluating its ability to provide answers with in- and out-of-domain passages. The second problem revolves around fine-tuning Roberta on the COVID-19 QA dataset and testing its capabilities. We test the performance based on Exact Match (EM) and F1 scores.

1 Introduction

In NLP, transformer-based models like RoBERTa, a variant of BERT, have demonstrated excellent performance in question-answering systems, especially on the SQuAD 2.0 dataset. In this assignment, we explore its ability to generalize to domain-specific and out-of-domain data to test it and look for improvements after fine-tuning it for specific datasets.

In this assignment, we evaluate the effectiveness of fine-tuning and Adapter-based training for enhancing the model's QA capabilities in specialized contexts.

2 Problem 1: Robustness in QA

RoBERTa performs impressively with well-organized, specialized datasets, but the real world throws some curveballs that can be more difficult to navigate. In this exploration, we're looking into how well RoBERTa can manage these unpredictable conditions. We'll test its ability to tackle various texts, including training, modified passages with specific questions, and completely different contexts. Our goal is to see how adaptable RoBERTa is across multiple scenarios.

2.1 In-domain Passages

For this part, evaluate RoBERTa's performance on in-domain passages sourced from Wikipedia for topics like - Game of Thrones, Youtube, Dark Matter, Bermuda Triangle, and Apple. These topics are vast and have much information available on the internet.

The first step was to scrape the wiki and get the info in under 5000 words, and each passage was saved as a text file.

Excerpts from passages:

- *It is an adaptation of A Song of Ice and Fire, a series of fantasy novels by George R. R. Martin, the first of which is A Game of Thrones. The show premiered on HBO in the United States on April 17, 2011, and concluded on May 19, 2019, with 73 episodes broadcast over eight seasons.*
- *These include modified Newtonian dynamics, tensor–vector–scalar gravity, or entropic gravity. So far, none of the proposed modified gravity theories can describe every piece of observational evidence simultaneously, suggesting that even if gravity has to be modified, some form of dark matter will still be required.*

The answers received were as follows:

- How many seasons were in Game of Thrones?: eight
- What is the final episode of Game of Thrones?: May 19, 2019
- Who are the founders of YouTube?: Steve Chen, Chad Hurley, and Jawed Karim
- How much did Google pay for YouTube in 2006?: $1.65 billion
- When was YouTube founded?: February 14, 2005

- What is YouTube's most popular video category?: music videos
- What is dark matter made of?: primordial black holes
- What percentage of the universe's mass is made up of dark matter?: 85%
- Is dark matter observable in the laboratory?: difficult to detect in the laboratory
- Where is the Bermuda Triangle located?: North Atlantic Ocean
- What is the Bermuda Triangle's nickname?: Devil's Triangle
- Is the Bermuda Triangle considered dangerous by experts?: reputable sources dismiss the idea that there is any mystery.
- How many apple cultivars exist?: more than 7,500
- Why are apple trees grafted onto rootstocks?: botanical evaluation
- What are the primary problems that apple trees face?: fungal, bacterial, and pest
- What are apples commonly grown for?: cooking, eating raw, and cider or apple juice production

While it could answer the above questions, below are the ones it couldn't:

- Who is the main character in Game of Thrones?
- What is dark matter made of?
- How does dark matter interact with light?
- Who discovered the first Apple?
- What percentage of the universe's mass comprises dark matter?
- Is dark matter observable in the laboratory?

2.2 In-domain passages Result

- **Strong on Explicit Facts:** Finds accurate information about various topics, like who started YouTube, significant historical dates, and important numbers.
- **Weak on Implicit Queries:** Struggles with questions that involve reasoning or are unclear (like asking about the main character in *Game of Thrones* or how dark matter interacts).
- **Inconsistent Across Domains:** It performs well in certain topics but doesn't do as well in others.

2.3 Modified Passages

In this part, we look into manipulating RoBERTa's responses by modifying the passages. We alter some key details and check how well it responds.

Example modifications:

- **Show Genre:** Changed from "fantasy drama" to "sci-fi drama".
- **Creators:** "David Benioff and D. B. Weiss" → "David Beanoff and D. B. Weasel".
- **Novel and Author:** A Song of Ice and Fire by "George R. R. Martin" → A Song of Suspenseful Novels by "George R. R. Martian".
- **Premiere and End Dates:** "April 17, 2011" – "May 19, 2019" → "April 1, 2012" – "May 1, 2020".
- **Episodes and Seasons:** "73 episodes, eight seasons" → "72 episodes, seven seasons".
- **Setting:** "Westeros and Essos" → "Westerland and Essy".
- **Iron Throne Arc:** "Seven Kingdoms of Westeros" → "Satan Kingdoms of Westerland".
- **Night's Watch:** Changed to "Night's Watchers".
- **Awards:** "59 Primetime Emmy Awards" → "59 Pizza Rolls Awards".
- **Prequels:** House of the Dragon (2022) and A Knight of the Seven Kingdoms (2025) → House of the Dragonballs (2023) and A Knight of the Seven Viking Kingdoms (2090).
- **Founders' Names:** Added nicknames for founders: • "Steve Chen" → "Steve the Master Chen" • "Chad Hurley" → "Chad 'I invented the Internet' Hurley" • "Jawed Karim" → "Jawed 'I'm the true genius' Karim".
- **User Stats Exaggeration:** "2.7 billion monthly active users" → "2.7 gazillion" "One billion hours watched daily" → "One trillion hours".
- **Cat Video References:** Videos uploaded at "500 hours of content per minute" → "500 hours of cat videos per minute". "Mostly about cats" was added to describe the 14 billion videos.
- **Google Acquisition:** "YouTube was purchased by Google" → "YouTube was kidnapped by Google".

Honorable Mentions:

- "Ruling dynasty" to "ruling watermelon".
- "Frying pan fan base" instead of "active fan base".
- "Violence (including bad violence)" and "massive controversy" added for extra chaos.
- Changed "greater audiences" to "mostly of teenagers who can't afford cable".
- Added sarcastic remark, "who needs safety anyway?", regarding endangering children's well-being.

Below are the questions RoBERTa answered:

- From sample1.txt
  - Who is the legendary mastermind behind Game of Thrones? → "David Benioff and D. B. Weiss"
  - How many seasons were in Game of Thrones? → "eight"
  - What is the final episode of Game of Thrones? → "May 19, 2019"
- From sample2.txt
  - Who are the legendary masterminds behind YouTube? → "Steve Chen, Chad Hurley, and Jawed Karim"
  - How much did Google splash out for YouTube in 2006? → $1.65 billion

Below are the questions that it did not answer:

- Who is the main character in Game of Thrones?
- What year did Game of Thrones begin filming?
- When did YouTube magically appear on the internet?
- What is YouTube's top-tier video category, aka the undisputed champion?

2.4 Modified passage Results

- RoBERTa struggled with more ambiguous or subjective questions, such as identifying the "main character" of Game of Thrones.
- It performed well with factual, numerical, and direct lookup questions, such as the number of seasons or YouTube's acquisition cost.

Table 1: Comparison of RoBERTa's Responses in Part 1 vs. 2

Aspect	Part 1	Part 2
Unanswered Questions	Fewer questions were unanswered	Many questions were left unanswered, even those previously answered correctly
Passage Rewording	Handled straightforward questions well	Struggled with reworded or exaggerated questions

2.5 Out of domain passage

For this part, the passage given was entirely out of domain. This is the passage that Roberta was never trained on. Thus, it has never seen this.

Example excerpt:

Harry looked into the fire. Now he came to think about it... every odd thing that had ever made his aunt and uncle furious with him had happened when he, Harry, had been upset or angry... chased by Dudley's gang, he had somehow found himself out of their reach... dreading going to school with that ridiculous haircut, he'd managed to make it grow back... and the very last time Dudley had hit him, hadn't he got his revenge, without even realizing he was doing it? Hadn't he set a boa constrictor on him? Harry looked back at Hagrid, smiling, and saw that Hagrid was positively beaming at him.

"See?" said Hagrid. "Harry Potter, not a wizard – you wait, you'll be famous at Hogwarts."

"The little bronze ones."

Harry counted out five little bronze coins, and the owl held out his leg so Harry could put the money into a small leather pouch tied to it. Then he flew off through the open window.

- **Struggle with Named Entities**
  - RoBERTa misidentified "Albus Dumbledore" as the main character instead of "Harry Potter."
  - It failed to recognize "Harry" and "The Boy Who Lived," which are synonymous in the Harry Potter universe.
- **Conceptual Definitions**
  - It did not answer "What is a muggle?" because the model struggles with abstract definitions outside factual retrieval.
  - Similarly, it left "Who is a wizard?" unanswered.
- **Performance on Explicit Locations & Titles**
  - Correctly identified "Privet Drive" as the residence of Mr. and Mrs. Dursley.
  - Recognized "Professor McGonagall" as a professor, which aligns with the text.

2.6 Out of domain result

Table 2: RoBERTa's Performance on Out-of-Domain Data

Question	RoBERTa's Answer
Who is the main character?	Albus Dumbledore (Incorrect)
Who is Harry?	(No Answer)
Who is the boy who lived?	(No Answer)
Who is a professor?	Professor McGonagall (Correct)
Where do Mr. and Mrs. Dursley live?	Privet Drive (Correct)
What is a muggle?	(No Answer)
Who is a wizard?	(No Answer)
Who lives in Godric's Hollow?	(No Answer)

3 Problem 1 Results

- Accurately retrieved factual answers like *Game of Thrones* creators and YouTube's acquisition price.
- Failed when reasoning was needed, e.g., identifying *Game of Thrones*' main character or *Harry Potter* synonyms.
- Couldn't define terms like *muggle* or *wizard*, showing reliance on phrase-matching over deeper understanding.
- Correctly retrieved structured details like *Privet Drive* and *Professor McGonagall*, favoring clear factual mentions.
- Misidentified key names (*Dumbledore* as the main character) and left many questions unanswered in *Harry Potter* passages.

4 Problem 2: Domain Adaptation for QA

For problem 2, we are exploring training in a question-answering model. In this part, we will use the COVID-19 Question Answering dataset and evaluate RoBERTa's performance on EM and F1 metrics. Next, we will fine-tune it, evaluate the performance after doing so, and compare it.

4.1 Dataset

The Covid-QA dataset contains 2,019 question-answer pairs.

Table 3: Dev Set

Question	Answer
What is Clostridium difficile?	Gram positive, anaerobic bacterium
What is sporulation?	Adaptive strategy that enables bacteria to survive harsh environmental conditions for prolonged periods of time
What is the key regulator to sporulation?	Spo0A

Table 4: Test Set

Question	Answer
What causes viral-induced exacerbations?	Viral infection increases airway inflammation.
What advances help study viral exacerbations?	3D cultures, organoid models, and challenge models.
What is a major source of airway disease exacerbation?	Respiratory virus infection.

Table 5: Train Set

Question	Answer
How much has translational research increased?	Exponentially, up 1800%.
What does the author call this dilemma?	'Information economy paradox'.
How do many viruses resolve this?	By manipulating host cell proteins.

4.2 Part 1

In this experiment, we evaluate the zero-shot performance of RoBERTa fine-tuned on SQuAD2 (pretrained model: deepest/Roberta-base-squad2) on the Covid-QA dataset. The performance of the model is measured using two key metrics: Exact Match (EM), The percentage of predictions that match the correct answer exactly, and F1 Score, The harmonic mean of precision and recall, which accounts for partial matches.

4.2.1 Observations

Below is how the model predicted the development and test set each.

Table 6: Predictions on Development Set

ID	Prediction
2988	enteric cytopathic human orphan virus
2989	respiratory illness, hand-foot-and-mouth disease
2990	California Department of Public Health
2991	within 1 day of identification of etiology
2992	CDPH
2993	E-30

Table 7: Predictions on Test Set

ID	Prediction
3864	complexity and heterogeneity of the disease
3865	impaired bacterial immune response
3866	interplay between virus and pathogenic bacteria
3867	morbidity and sometimes mortality in patients
3869	respiratory viral infection
3870	infiltration of activated immune cells

4.2.2 Results

As the results below suggest, we have a low EM score, indicating that the model did not extract the exact answers. The F1 score for both development and test sets is around 40%. This shows that the model does capture information, but it is not always correct. We can also see that the test performance is worse than the development set.

Table 8: Evaluation Results

Metric	Dev Set	Test Set
Exact Match (EM)	27.09%	24.27%
F1 Score	46.38%	43.79%
Total Samples	203	375
Has Answer EM	27.09%	24.27%
Has Answer F1	46.38%	43.79%
Has Answer Total	203	375

4.3 Part 2

The training setup is based on the RoBERTa model to be fine-tuned on the COVID QA dataset, following a structured approach using the HuggingFace's Trainer class for training and evaluation. However, as fine-tuning was not explicitly required, the model was not actually trained, but the framework was fully prepared for execution.

For dataset preprocessing, the context questions are tokenized to a maximum length, and the answers are aligned from start to end to map them to the tokenized input.

Below are the fine-tuning parameters:

Table 9: Fine-Tuning Configuration for RoBERTa on Covid-QA

Parameter	Value
Model Checkpoint	`deepset/roberta-base-squad2`
Tokenizer	RoBERTa Tokenizer
Maximum Sequence Length	384
Learning Rate	2×10^{-5}
Batch Size (Train/Eval)	8
Number of Epochs	3
Weight Decay	0.01
Evaluation Strategy	End of each epoch
Save Strategy	Best model based on F1-score
Gradient Clipping	1.0
Mixed Precision (fp16)	Disabled
Logging Steps	500
Checkpoint Limit	2

4.4 Part 3

As discussed in the part above, tuning an entire transformer model for Question answering is not exactly efficient. Thus, instead of fine-tuning the model, we use adapters.

Adapter-based fine-tuning reduces computational overheads and maintains good performance.

But, for this part, we use LoRA (Low-Rank Adaptation) finetuning by using 'peft' and 'transformers' libraries. Using LoRA, we introduce low-rank updates to the attention layers to reduce memory usage and training time. After defining hyperparameters such as learning rate, batch size, and epochs, we train the model and monitor it through early stopping to prevent overfitting. Finally, predictions are generated for the development and test sets and saved as JSON files for further evaluation.

Below are the hyperparameters finalised after getting the best results:

Table 10: Hyperparameters for Model Training

Parameter	Value
Model Checkpoint	`deepset/roberta-base-squad2`
Tokenizer	RoBERTa Tokenizer
Learning Rate	0.001
Number of Training Epochs	5
Weight Decay	0.1
Per Device Train Batch Size	32
Per Device Eval Batch Size	8
Number of Warmup Steps	1000
Logging Steps	15
Evaluation Steps	15
Save Strategy	Epoch
Load Best Model at End	True
Save Total Limit	3
Gradient Accumulation Steps	2
Gradient Checkpointing	True
Metric for Best Model	eval_loss
LoRA Rank	8
LoRA Alpha	32
LoRA Dropout	0.1
LoRA Target Modules	`[query, value]`

Below is the summary of the best run:

Table 11: Summary of training metrics for each epoch

Epoch	Loss	Grad Norm	Learning Rate
1	13.3165	18.32	4.67e-06
2	13.3226	19.23	4.33e-06
3	13.3071	18.98	4.00e-06
4	13.2241	19.07	3.67e-06

4.4.1 Observations

Below is how the model performed on the development and test set.

Table 12: Predictions on Development Set

ID	Prediction
430	nine
916	clostridial toxins A and B
5287	The capacities of nanopore sequencing for viral diagnostics
5288	19.2 to 103.5X
5289	lower than 7

Table 13: Predictions on Test Set

ID	Prediction
3864	complexity and heterogeneity of the disease
3865	episodic exacerbations of the disease
3867	morbidity and sometimes mortality in patients
3872	high
3873	ease of transmission and infection

4.4.2 Results

As the results suggest, the model performed better on the development set with an F1 of 63 and slightly lower on the test set.

Table 14: Evaluation Results

Metric	Dev Set	Test Set
Exact Match (EM)	38.42%	30.13%
F1 Score	63.99%	61.16%
Total Samples	203	375
Has Answer EM	38.42%	30.13%
Has Answer F1	63.99%	61.16%
Has Answer Total	203	375

5 Problem 2 Results

- The initial RoBERTa model, without fine-tuning, struggled with low Exact Match (EM) scores: 27.09% on the development set and 24.27% on the test set.
- The F1 score for the zero-shot model was better, with 63.99% on the development set and 61.16% on the test set, indicating that while it didn't always get the exact answer, it still captured useful information.
- After applying LoRA, the model showed improvement: the EM score rose to 38.42% on the dev set and 30.13% on the test set, meaning it got closer to the correct answers.
- F1 scores saw a noticeable boost on the development and test sets, reflecting better overall accuracy.
- In summary, LoRA fine-tuning helped the model perform better for the Covid-QA dataset.

6 Conclusion

- **In-Domain Performance:** RoBERTa shows strong performance when answering straightforward, factual questions, such as those that ask for concrete information (e.g., "How many seasons are there in Game of Thrones?"). However, it struggles with more abstract or subjective queries, such as identifying the main character in Game of Thrones. This suggests that while RoBERTa excels in well-defined domains, it may falter when deeper reasoning or interpretation is required.
- **Modified Passage Performance:** When the questions or passages were slightly altered, RoBERTa's performance declined. It struggled with rephrased or exaggerated questions, indicating that the model may rely more on surface-level phrase matching rather than a deeper understanding of the question. However, it still performed well when factual questions were phrased similarly to the original dataset.
- **Out-of-Domain Challenges:** RoBERTa faced challenges with out-of-domain data. For example, it misidentified key characters in the Harry Potter series and failed to correctly define terms like "muggle" or "wizard." This highlights RoBERTa's limitations when dealing with data that deviates from the training domain.
- **Domain Adaptation:** RoBERTa's performance on the Covid-QA dataset, when tested without fine-tuning, was suboptimal. Its Exact Match (EM) and F1 scores were low, indicating that RoBERTa requires domain-specific fine-tuning to achieve better accuracy. However, the model still managed to capture some relevant information, albeit with lower precision.
- **Adapter-based Fine-Tuning:** The use of LoRA for adapter-based fine-tuning proved to be an efficient approach. This method allowed RoBERTa to be adapted to the Covid-QA dataset without the computational burden of full model fine-tuning. Although the results were improved, there is still room for further enhancement in terms of precision and performance.
- **Overall Insights:** This study reveals that RoBERTa is effective for answering well-defined, factual questions but struggles with complex reasoning, abstract concepts, and out-of-domain applications. Fine-tuning, especially with adapter-based methods like LoRA, enhances RoBERTa's ability to handle specialized tasks. However, the results underline the need for additional improvements in handling complex queries, domain generalization, and reasoning tasks.