“Please Answer the Question” - Applying a QNLI Approach to Interview Settings

Andrew Wright

Abstract

The field of natural language processing has made considerable improvements over the past decade towards question answering, ranging from generating text that answers a given question to extracting text from within a given source that answers the question. This has a wide range of applications, including providing researchers a better understanding of what it means to answer a question from a machine’s point of view. Our aim is to use this understanding and specific methods from the field of question answering and natural language inference to evaluate responses to questions from humans.

Introduction

There are many real-world settings where people fail to answer questions given to them, whether due to their lack of knowledge of the answer, their inability to answer, or purposeful evasion. Our goal is to produce a fair and accurate tool for determining whether a person has actually answered a given question in an interview or similar format. A common theme, particularly in political settings, is to blame a respondent for “dodging” the questions - this paper aims to create a model that can provide insight into when and where that claim is true. This is both an interesting and difficult task because it is often not obvious, and many public facing individuals are good at “hiding” to a general audience that they are not really answering a question.

Background

Significant and consistently improving advancements have been made in recent years across the natural language processing field as a whole, as well as in its subfields most related to this problem at hand. We will leverage these advancements to choose the data and models that will have the best chance at successfully performing our desired task.

A foundational paper for large-scale reading comprehension models is SQuAD[1], which created a large dataset of natural language question-answer pairs from Wikipedia and trains a relatively basic model to find the correct answer to a given question within a passage. From here, the GLUE QNLI dataset converts SQuAD into sentence pairs with the task of determining whether the first (question) entails the second (answer)[2]. This task naturally fits what we are trying to achieve here, and will be a main starting point for our work. The dataset also has some particularly useful qualities which we will take advantage of. It specifically includes cases where the most similar sentence to the question was NOT the answer. This should help our model

[1] https://aclanthology.org/D16-1264/
[2] https://aclanthology.org/W18-5446/

learn cases where the respondent attempts to evade a question by answering with something relevant to the topic, but that does not actually answer the question.

Still, the QNLI dataset is not ideal for this task on its own. It is limited in that the question-answer pairs tend to be relatively simple, and may not capture the nuance and complexity of the types of questions and answers found in in-depth interviews. The QuAC [3] and CoQA[4] datasets are both more conversational datasets that include longer-form questions and answers. We expect these to be closer in nature to a more complex interview setting, and therefore training on these datasets may prove beneficial for learning this task.

We will also follow the latest advancements in transformer based encoder-only models. The base model we plan on using for this task is ALBERT [5], which follows a similar architecture but improves on the ubiquitous BERT model on many relevant tasks including GLUE QNLI. It also importantly provides efficiencies by reducing parameters in order to make the model quicker and less resource-intensive to train, which will be important given the limited time and resources. It is common practice in the field to fine-tune BERT and similar models for specific classification tasks[6], which we will be attempting to do here.

Methods

We hope that a model trained on this task will generalize well to the use case of interview settings. The initial data we chose to measure against are transcripts from the New York Editorial Board's interviews for the mayor of New York City in 2025. This dataset is relevant to the current time, is relatively clean, and has nuanced conversations with numerous cases of pushback and back-and-forth discussions on the same topics, making it a strong starting point for this task.

This dataset contains a total of 458 question-answer pairs. I preprocessed the data into JSON format with each element containing the question, answer, follow-up question, interviewer, and interviewee.

A novel approach was used to label the data. I leveraged 3 large language models: GPT-4o, Claude-3-7-sonnet, and Gemini 2.0 to label the data, and used a committee voting algorithm to choose the final label in the case of disagreement. The label itself is meant to be a proxy for whether or not the respondent answered the question. Specifically, I provided each LLM with the asked question, as well as the follow-up question after the response to the initial question, and asked it to label "1" for cases where the follow-up repeats the initial question or states that the question was not answered. This is meant to be a more objective and straightforward task than asking the LLM to actually label whether or not the question was answered.

[3] https://arxiv.org/abs/1808.07036
[4] https://aclanthology.org/Q19-1016/
[5] https://arxiv.org/pdf/1909.11942v6
[6] https://www.sciencedirect.com/science/article/abs/pii/S0172219019300742

I began the labeling process by taking a selection of 10 data points, and self-labeling them. The specific data points were chosen to be a mix of easy and difficult problems, with relatively unambiguous labels. From here, I asked each LLM to provide their own label based on the criteria explained above. I calibrated the prompts slightly for each model to align with the self-labeled responses (appendix 1) - although they were equivalent in all but one case before this calibration. This was an important validation step to ensure correctness of the labels. Of course, this process is still not perfect, and in an ideal world we would have used a fully manually labeled or at least manually reviewed dataset. Appendix 5 contains the final prompts and system messages used. From here, I used these calibrated prompts to label the full dataset.

I arrived at the committee voting algorithm by noting that Gemini and Claude had the most similar responses across the dataset, and from a small random selection of scored question-response pairs, aligned most closely with self-determined labels. I took the label from these two models in cases where they agreed, and used GPT-4o as a tie-breaker in the cases where they disagreed. Figure 1 below shows the agreement across these models' labels, counting the number of question-answer pairs for which they agreed.

Figure 1 - LLM Label Agreement

	GPT	Claude	Gemini
GPT	458	412	422
Claude	-	458	432
Gemini	-	-	458

A supplementary dataset called INTERVIEW[7] was added later on in the process. This dataset contains interview data from NPR broadcasts over 20 years from 1999 to 2019. The purpose of this dataset is to be a large-scale source that models real world conversations. This has pros and cons compared to the Editorial Board data. Its large size seemingly solves the problem of not having sufficient data, but the nature of the data itself is not quite as conducive for the problem at hand. While we took the same approach to labeling, there were significantly fewer instances of "1" labels. The interviews have a more casual tone, and the interviewer is less likely to push back even when a valid or sufficient response to the question was not provided. While the dataset contains over 100,000 rows of data, we filtered this down to a much more manageable, but still useful size of 20,000. Appendix 2 contains a summary of the labels for this dataset in the same manner as figure 1.

For the interview task, a simple baseline is used as the objective we are attempting to beat. The natural baseline for imbalanced data such as this is predicting all 0s - or that all questions are

[7] https://arxiv.org/pdf/2004.03090

being answered sufficiently. We see that this of course has high accuracy, but zero recall and precision. Outside of the extreme edge cases that would (predicting all 1s for example) we hope to see an improvement in recall with little to no loss of accuracy and precision in our final model. This would indicate that the model is actually learning to predict the target class. While a bag of words or similar approach could lead to a more sophisticated baseline, this task is complex enough where there are not obvious choices that would be a natural fit for these types of baseline models.

Results and discussion

The training on QNLI and CoQA was relatively straightforward, although CoQA has a slightly tricky data structure to work with. The bulk of the experimentation here was done on the interview task, which the following section will walk through.

Experiment 1.1 serves as the simplest iteration beyond the baseline, effectively acting as a secondary baseline for subsequent experiments. The goal of this was to ideally beat the initial baseline and provide a starting point for further improvements. A weighted cross-entropy loss was used for backpropagation in order to encourage the model to label “1”s, since the model predicted all “0”s otherwise due to the large imbalance. For model validation, due to the limited size of the dataset we used 4-fold cross validation for initial validation and adjusting hyperparameters, and then trained a final model on the full set of training data to be used for test set predictions.

Figure 2 - Test Set Results by Experiment

	Accuracy	Precision	Recall	f1
Baseline	0.9275	0	0	0
ALBERT - QNLI only (1.1)	0.65217	0	0	0
ALBERT - QNLI + CoQA (2.1)	0.89855	0	0	0
ALBERT - CoQA+ QNL (2.2)	0.88406	0	0	0
ALBERT - CoQA+ QNLI + INTERVIEW (3.1)	0.88465	0.09119	0.30851	0.14077

Figure 2 outlines the test set results for this experiment as well as subsequent ones. This initial experiment did not prove to be successful. At this point there were a couple of possible reasons for this suggesting different directions to go in for improvement. The first of these was to try to improve the model itself. Drawing from the learnings of the T5[8] paper as well as subsequent related work, I aimed to leverage transfer learning to strategically train the model to be better prepared for this task. As discussed, the QNLI dataset does not contain the nuanced and conversational content that is contained in these interviews. As such, I chose to train the model on the CoQA dataset in order to attempt to learn more of this nuanced conversation that will be required for the task. We know from existing literature that there may be different transfer learning benefits depending on the order in which we train on the different datasets, so I ran experiments training in both orders (experiments 2.1 and 2.2 respectively).

Although there were slight improvements in both of these models in the validation data (see appendix 3), they were once again unsuccessful on the test set. Experiment 2.2, training on CoQA first and then QNLI, offered marginally better results on the validation data compared to 2.1. After running these experiments it became more clear that the issue had more to do with the data, and particularly the quantity of it, versus the model. Specifically, based on these results it seems like the model was in fact learning the training data, but was overfitting to the specific cases with label 1. The imbalanced nature of the dataset made this problem worse.

A closer examination of the cases that were falsely labeled as 1 in experiment 2.2 (see appendix 4) showed that they were not completely unreasonable guesses by the model. The first 2 are cases where the question is answered, but where the response actually cuts off the question partway through. This could indicate a lack of coherence between the question and response which the model could be picking up on. This suggests that the model is still learning something relevant, but is not yet accurate or precise enough to provide consistent results.

Experiment 3.1 added the supplementary INTERVIEW dataset, labeled and processed in the same way as the original dataset. The interview dataset significantly expanded the amount of data, and importantly the number of positive labels.

Given the results of this experiment, I attempted numerous variations of the fine-tuning process in order to get this to work. These included altering which layers are being trained using the following combinations: classifier only , classifier and pooler, classifier, pooler, and feed forward layer. This was meant to help prevent the model from overfitting embedding layers to the dataset, but these approaches did not make notable improvements to the model based on validation results.

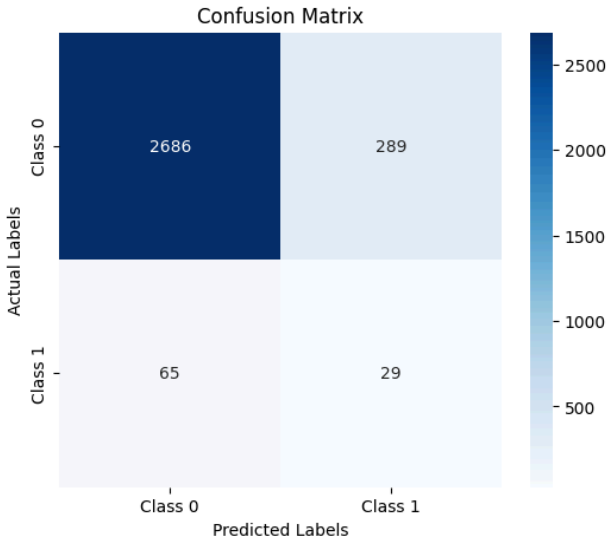
In order to tackle the problem of imbalance data, techniques from Cost-Sensitive BERT[9] along with methods covered here[10] were used as inspiration. Undersampling the majority class is probably the most simple of these methods, and was tried for experiments 2.2 and 3.1, but still

[8] https://arxiv.org/pdf/1910.10683
[9] https://arxiv.org/pdf/2003.11563
[10] https://arxiv.org/pdf/1710.05381

led to overfitting with very low model precision on the validation set. Cost-sensitive learning, or weighting the minority (positive) class to carry much more loss for the Adam optimizer, is ultimately what proved most successful here. While the natural values for label weights for these types of problems is the inverse proportion of their incidence, this still led to the model predicting all 0s in most cases. Adjusting the weights to be even more extreme led to a better outcome that is listed in the results table for experiment 3.1.

These results ended up being a mixed bag. In terms of the stated goal of beating the baseline, this final model did succeed. While the raw accuracy score is slightly lower than the baseline, the f1 score is higher than could be gotten by random predictions on average. If we were to naively select the same number of positive labels (318) at random as this model scored on the test set, we would expect to get a similar precision but much lower recall. We can therefore say that this model has made a substantial improvement over the baseline at predicting the positive class. On the other hand, the results cannot be considered great. Even the recall here is not high compared to the results of other successful fine-tuning experiments in the field.

Figure 3 - Confusion matrix of test set predictions (Experiment 3.1)



Conclusion

While we were able to get this model to perform slightly better than the baseline, indicating some level of learning the task, more work will be needed to get it to the point where it is accurate and ultimately useful. This problem is particularly challenging because the nature of the task leads to very nuanced differences between the two outcome classes in many cases, and is exacerbated by the relative rarity of positive labels, leading to imbalance data . These together both make it difficult to come up with a simple approach such as a bag of words, and also means that a more in-depth approach such as the one here will likely need significantly more data and higher quality labels. With additional time and resources, employing oversampling techniques as used in the Cost-Sensitive BERT paper could potentially improve

the model. A large-scale curated and more balanced dataset would also likely be needed for higher quality outcomes here. Other modeling approaches such as few-shot learning with state of the art LLMs on their own could also prove successful.

Appendix

Appendix 1 - Manual Label Validation Process
https://github.com/AndGWright/interview-responses/blob/main/llm_label_example.ipynb

Appendix 2 - NPR Dataset Labels

	GPT	Claude	Gemini
GPT	20000	17867	17929
Claude	-	20000	19420
Gemini	-	-	20000

Appendix 3 - Validation Set Outcomes
Note: The baseline used for QNLI is a simple random forest aimed to validate that the ALBERT models are effectively learning this task. The table shows the results for the best experiments for each type and is meant to demonstrate learning of the training data, but overfitting to this data, up until experiment 3.1.

	GLUE QNLI	Interview Task - Accuracy	Interview Task - Precision	Interview Task - Recall	Interview Task - f1
Baseline	0.597	0.9275	0	0	0
ALBERT - QNLI only	0.833	0.896907	0.285714	0.285714	0.285714
ALBERT - QNLI + CoQA	0.833	0.938144	1	0.142857	0.25
ALBERT - CoQA+ QNLI	0.835	0.918367	0.400000	0.285714	0.333333
ALBERT - CoQA+ QNLI + supplementary dataset	0.835	0.894212	0.101626	0.312500	0.153374

Appendix 4 - Experiment 2.2 Misclassified Positive Labels

Experiment 2.2 False Positives	
question	response
But is a loud, raucous protest with posters that say “Israel equals genocide” and things of that nature—	“From the river to the sea.” “Globalize the Intifada.”
Follow-up transit question: you said the biggest MTA problem is the lack of investment —	The historical one.
You forgot something on Adams that you appreciate.	The NYC Reads program.

Appendix 5 - LLM Labeling Prompts

Claude-3-7

Prompt:
“You are an expert in natural language processing. Please label the following question-follow-up pair as "1" if the follow-up asks the same or a similar question to the initial question, or if it explicitly or implicitly states that the question was not answered. Otherwise, label "0". It is okay to label 0 if the follow-up is asking for further clarification or specifics.”

System message:
“You analyze question-follow-up pairs and determine if the follow-up is asking the same/similar question or indicating the initial question wasn't answered (label '1'), or if it's a new topic or asking for further details (label '0'). Respond with ONLY the label (1 or 0)."

GPT-4o

Prompt:
“Label the following question-follow-up pair as '1' if the follow-up repeats the question or suggests the question was not answered. Otherwise, label '0'.\n\n. It is okay to label 0 if the follow-up is asking for further clarification or specifics, or if the follow-up is unrelated."

System message:
"You are an NLP model trained to label question-follow-up pairs."

Gemini 2.0

Prompt:
“Label the following question-follow-up pair as '1' if the follow-up repeats the question or suggests the question was not answered. Otherwise, label '0'.\n\n. It is okay to label 0 if the follow-up is asking for further clarification or specifics, or if the follow-up is unrelated."