Tanisha Khurana
Student ID - 200483139

CSC 595+791 Natural Language Processing, Fall 2023
Assignment Project -2

Introduction

The main task for this assignment is understanding Large Language Models (LLMs) and exploring their capabilities, limitations, and ethical considerations. To attain a comprehensive understanding on how to harness the power of LLMs effectively.

In the pursuit of comprehending LLM's and unlocking their full potential, prompt engineering emerges as a critical element. My exploration specifically delved into deploying LLMs for a question answering application, where the intricate design of prompts played a central role.

Prompt engineering, in the context of question answering, entails crafting input instructions that prompt the LLM to provide accurate, relevant, and contextually appropriate responses. The objective is not merely to generate grammatically correct questions but to understand the idiosyncrasies of the LLM's response patterns. This involves experimenting with different phrasings, structuring questions to provide more context, or introducing specific cues to guide the model toward the desired outcome. In my experimentation, I employed various prompt strategies to gauge their impact on the question answering performance of the LLM.

For this Question Answering task, I am using SQUAD2.0 dataset. Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

For this task, I experimented with F1 metric score, Exact Match (EM) score, Bi-Encoder Score, Semantic Similarity or Semantic Answer Similarity (SAS) metric, Bilingual Evaluation Understudy (BLEU), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score.

Exact Match: This metric is as simple as it sounds. For each question+answer pair, if the characters of the model's prediction exactly match the characters of (one of) the True Answer(s), EM = 1, otherwise EM = 0. This is a strict all-or-nothing metric; being off by a single character results in a score of 0. When assessing against a negative example, if the model predicts any text at all, it automatically receives a 0 for that example.

F1: F1 score is a common metric for classification problems, and widely used in QA. It is appropriate when we care equally about precision and recall. In this case, it's computed over the individual words in the prediction against those in the True Answer. The number of shared words between the prediction and the truth is the basis of the F1 score: precision is the ratio of the number of shared words to the total number of words in the prediction, and recall is the ratio of the number of shared words to the total number of words in the ground truth.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

BLEU Score:

BLEU is a metric commonly employed to measure the similarity between a generated text and one or more reference texts. Originally designed for evaluating machine translation, BLEU calculates precision by comparing the n-grams (contiguous sequences of n items, usually words) in the generated text to those in the reference text. The score ranges from 0 to 1, with 1 indicating perfect similarity. BLEU considers precision at various n-gram levels, typically from unigrams (individual words) to n-grams of four or more words.

ROUGE Score:

ROUGE is a set of metrics designed to evaluate the quality of summaries by comparing them to reference summaries. The key idea behind ROUGE is to assess the overlap of n-grams and word sequences between the generated and reference summaries. Like BLEU, ROUGE measures precision, recall, and F1 score. ROUGE is particularly valuable for tasks like text summarization, where capturing the essence of the content is crucial.

ROUGE has several variants, including ROUGE-N (measuring overlap of n-grams), ROUGE-L (measuring longest common subsequence), and ROUGE-W (measuring word overlap). These metrics provide a more comprehensive evaluation of the generated text's quality, taking into account both precision and recall.

Other metrics such as Bi-Encoder Score, Semantic Similarity or Semantic Answer Similarity (SAS) metric were also used.

Bi-Encoder Score is based on sentence transformers architecture. It uses two language models to separately calculate embeddings for predicted answers and answer labels. Later cosine similarity is used to calculate the score between contextual embeddings. Before calculating embeddings, two language models are trained on the multi-lingual paraphrase dataset and STS

benchmark dataset. One of the advantages of the bi-encoder architecture is that the embeddings of the two text inputs (predicted answers and answer labels) are calculated separately.

Semantic Similarity or Semantic Answer Similarity (SAS) uses the "cross-encoder/stsb-roberta-large" language model, which has been trained on the STS benchmark dataset. Unlike Bi-Encoder where two separate models are used, SAS uses a cross-encoder architecture where a predicted answer and a label are separated by a special token to calculate the score. Among all neural-based metrics, cross-encoder model metrics have relatively the strongest correlation with human judgment.

Semantic Answer Similarity is often favored over traditional metrics like F1 score (which combines precision and recall) or Exact Match (EM) in question answering tasks due to its capacity to capture the nuanced semantic equivalence between answers. While F1 and EM emphasize word overlap and exact matching, they may overlook variations in phrasing, synonyms, or paraphrasing that are common in natural language. Semantic Answer Similarity, on the other hand, leverages advanced techniques such as cosine similarity with embeddings or contextualized representations to assess the broader meaning and context of answers. Semantic Answer Similarity metrics provide a more comprehensive and flexible evaluation, enabling models to be rewarded for conveying accurate information even if the wording differs, ultimately promoting a more robust and user-centric approach to question answering system evaluation.

Related Work on Squad v2 Dataset:

The Squad v2 dataset has been a focal point for the evaluation of question answering systems, prompting researchers to employ various approaches to tackle its challenges. One prevalent strategy involves the utilization of pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and its derivatives, as a foundation for QA models. These models demonstrate an adept understanding of context and have proven effective in extracting relevant information from passages.

Another common approach involves the use of ensemble models, where predictions from multiple models are combined to enhance overall performance. Ensemble methods often leverage diverse architectures or utilize different pre-training strategies to create a more robust and accurate QA system.

Additionally, fine-tuning techniques specific to the Squad v2 dataset have gained prominence. Researchers employ meticulous fine-tuning procedures to adapt pre-trained models to the unique characteristics of Squad v2, considering its inclusion of unanswerable questions. This involves

adjusting hyperparameters, incorporating adversarial training, or experimenting with domain-specific pre-training.

The current SOTA performance for the Squadv2 dataset is IE-Net(ensemble) score of 90.939 for EM and 93.214 for F1. FPNet (ensemble) score of 90.871 for EM and 93.183 F1.
IE-NetV2 (ensemble) score of 90.860 for Exact match and 93.100 for F1 score.

I have experimented with 3 prompting strategies namely:

1. **Zero-Shot Prompting**: Zero-Shot Prompting in prompt engineering tasks involves training models to perform specific tasks without exposure to any explicit examples during training. Instead, models are expected to rely on their pre-trained knowledge and generalization abilities to respond to prompts in new domains or tasks. This approach underscores the adaptability of language models, allowing them to infer patterns and information from context, making it particularly useful for handling diverse natural language processing tasks without the need for extensive task-specific training data.

2. **Few-Shot Prompting**: Few-Shot Prompting strikes a balance between zero-shot learning and fully supervised learning. In this method, models are provided with a limited number of examples or prompts during training, allowing them to learn from a small amount of task-specific data. This approach is advantageous when labeled data is scarce, as it enables models to leverage their pre-training knowledge while adapting to specific tasks through exposure to a handful of examples. Few-Shot Prompting showcases the ability of language models to extrapolate information and adapt to novel prompts with only a modest amount of task-specific guidance.

3. **Generated Knowledge Prompting**: Generated Knowledge Prompting involves dynamically incorporating external knowledge or information into prompts to enhance the performance of language models. This approach utilizes resources like knowledge graphs, databases, or information retrieval systems to construct prompts that augment the model's understanding. By integrating external knowledge sources into prompts, models gain access to a broader range of information, enabling them to generate more informed and contextually relevant responses. Generated Knowledge Prompting is particularly beneficial in scenarios where models need to demonstrate a deeper understanding of a topic by incorporating up-to-date or specialized information beyond their pre-training data, resulting in more accurate and contextually rich outputs.

   Results and Analysis:

   For prompting my gpt 3.5 model I used the following parameters:

Tanisha Khurana
Student ID - 200483139

Temperature is a measure of how often the model outputs a less likely token. The higher the temperature, the more random (and usually creative) the output. This, however, is not the same as “truthfulness”. For most factual use cases such as data extraction, and truthful Q&A, the temperature of 0 is best. Thus, for a QA model I used a temperature of 0.

max_tokens (maximum length) - Does not control the length of the output, but a hard cutoff limit for token generation. Ideally you won’t hit this limit often, as your model will stop either when it thinks it’s finished, or when it hits a stop sequence you defined.
I used max_token of 30.

Within the OpenAI API, messages often adopt specific roles to guide the model’s responses. Commonly used roles include “system,” “user,” and “assistant.” The “system” provides high-level instructions, the “user” presents queries or prompts, and the “assistant” is the model’s response. By differentiating these roles, we can set the context and direct the conversation efficiently.

For my initial task I used 10 samples each for zero shot and generated knowledge prompting and 20 samples were used for few shot where 10 were given with answers and 10 were the test cases.

10 samples result:
For zero shot strategy my average EM score is 30 i.e 30% of my output was an exact match to the ground truth and my mean average F1 score is 41.4166666.
My ROUGE score is 0.409147 for zero shot.
Other metrics I used were SAS which for zero prompt gave an average score of 0.49774858.

For the next prompt strategy I gave the question with an answer, telling the gpt on how to answer the question. Thus, forcing the gpt to reply exactly how the ground truth is replied to.

[[{'role': 'system',
'content': 'You are an AI assistant to answer questions. Please use your own knowledge to answer the questions. Please give very short answers. If you do not know the answer, please guess a most probable answer.'},
{'role': 'user', 'content': 'When did Beyonce start becoming popular?'},
{'role': 'assistant', 'content': 'in the late 1990s'},
{'role': 'user',
'content': 'What areas did Beyonce compete in when she was growing up?'}],

{'id': '56d43c5f2ccc5a1400d830aa',
'question': "What role did Beyoncé have in Destiny's Child?",
'model_answer': 'Lead vocalist',
'gt_answer': {'text': ['lead singer'], 'answer_start': [290]}},

Thus, I was able to prompt the model to give me brief answers just like the ground truth, increasing my EM and F1 score by a big margin.

{'exact': 100.0,
'f1': 100.0,
'total': 1,
'HasAns_exact': 100.0,
'HasAns_f1': 100.0,
'HasAns_total': 1,
'best_exact': 100.0,
'best_exact_thresh': 0.0,
'best_f1': 100.0,
'best_f1_thresh': 0.0},

The Knowledge Prompting gave the worst results since probably after giving the context, gpt gave longer answers decreasing metric score for all the defined metrics.

Metric	Zero-shot prompt	Few shot prompt	Generated Knowledge Prompting
F1	41.4166	75	24.706
Exact match	30	60	0
SAS	0.482	0.8473184	0.349
ROUGE	0.4091	0.75	0.75

Thus, few shot is currently the best prompting strategy for Question Answering applications giving the highest metric for SAS, F1 and EM.

100 samples results:

Tanisha Khurana
Student ID - 200483139

I took 100 samples for zero shot and knowledge prompt and 400 samples i.e 200 samples for few shot.

Metric	Zero-shot prompt	Few shot prompt	Generated Knowledge Prompting
F1	26.46388	48.80	25.5079
Exact match	16	36	0
SAS	0.48296937	0.52	0.38
ROUGE	0.2747	0.50	0.23

My F1 average score for zero shot gave me 26.8 and my EM score was 16, that means 16/100 times the answer was an exact match. My sas score remained almost the same. For few shot however my score increased drastically. 72 times out of 200 it was an exact match.

Error Analysis:
For zero shot prompting, I just gave a system prompt = "You are an AI assistant to answer questions. Please use your own knowledge to answer the questions. If you do not know the answer, please guess the most probable answer."
And then followed by the question. However since the SQUAD dataset is made to be used with the context as a sort of reading comprehension based application such questions like '**In what R&B group was she the lead singer?',** is ambiguous for the model in zero shot prompting.
'model_answer': "I'm sorry, but I don't have enough information to answer your question. Could you please provide the name of the person you are referring to?",
'gt_answer': {'text': ["Destiny's Child"], 'answer_start': [320]}},

Another qualm in evaluation metric for QA system is that F1 and exact match (EM) are not the correct metric for evaluating answers where even though the answer is the same, the language used for the answer is different, thus penalizing the actually correct answer.

predictions = [{'prediction_text': "**Beyonce started becoming popular in the late 1990s as a member of the group Destiny's Child**."]
references = [{'answers': { [**'in the late 1990s'**]}]

Even though we can see the answer is correct, my Exact match score is 0 and F1 is 37.5.

BLEU metric is not useful here as it's made for mostly machine translation tasks and it penalizes the model if my predictions are in uppercase and ground truth in lower case.

As we can see, even the addition of a period at the end decreases the score.

results = bleu.compute(predictions=['singing and dancing.'], references=['singing and dancing'])
{'bleu': 0.0, 'precisions': [0.75, 0.6666666666666666, 0.5, 0.0], 'brevity_penalty': 1.0, 'length_ratio': 1.3333333333333333, 'translation_length': 4, 'reference_length': 3}

results = bleu.compute(predictions=['Singing and dancing.'], references=["Singing and dancing."])
{'bleu': 1.0, 'precisions': [1.0, 1.0, 1.0, 1.0], 'brevity_penalty': 1.0, 'length_ratio': 1.0, 'translation_length': 4, 'reference_length': 4}

Thus, my average BLEU score was very low 0.01431, not a correct metric for such evaluations.

SAS metric also has a limitation as it also gives an incorrect answer a high score even though the answer is wrong.

Here are some examples of where my model went wrong:

Zero shot:
['there is no known connection between beyoncé and any acadian leader.'] ['joseph broussard.']
['christianity.'] ['methodist']
['houston, texas.'] ['fredericksburg']
["i'm sorry, i don't have access to personal information about beyonce's childhood."] ['imagine']

['mathew knowles.'] ['arne frager']
['columbia records.'] ['sony music']
['beyonce did not record with anyone for the movie "the best man."'] ['marc nelson']
['1993'] ['1996']
["the name destiny's child was not based on a quote in the bible."] ['book of isaiah']
['it is not publicly known who helped beyonce the most in fighting her depression.'] ['her mother']
['michelle williams and farrah franklin.'] ['farrah franklin and michelle williams.']
['mathew knowles'] ['beyoncé']
['jay-z'] ['her mother']

The reason might be that in the dataset there was a paragraph which gave context for such personal questions which the model gpt 3.5 is not trained on.

['lead vocalist.'] ['lead singer']
['"dangerously in love"'] ['dangerously in love']
['beyoncé released dangerously in love in 2003.'] ['2003']
['beyoncé won five grammy awards for her first solo album.'] ['five']

These are correct but their wordings are different.

Few shot:
'question': 'What song did she release in 2013 in response to a speech?',
'model_answer': '"Pretty Hurts"',
'gt_answer': {'text': ['Flawless'], 'answer_start': [445]}},

{'id': '56bfb502a10cfb140055125b',
'question': 'Between 2008 and 2009, which entertainers did Beyonce beat in earnings?',
'model_answer': "I don't have access to specific earnings data for that time period, but Beyonce has consistently been one of the highest-earning entertainers in the",
'gt_answer': {'text': ['Madonna and Celine Dion'], 'answer_start': [248]}}]

'question': 'Beyonce sang which song during the 2009 presidential inauguration?',
'model_answer': '"At Last"',
'gt_answer': {'text': ['America the Beautiful'], 'answer_start': [112]}},

'question': 'Beyonce described what as the "hardest thing she had to endure"?',

'model_answer': 'Childbirth',
'gt_answer': {'text': ['miscarriage'], 'answer_start': [617]}},

'question': 'When did Beyoncé release her fifth studio album?',
'model_answer': '2013',
'gt_answer': {'text': ['December 13, 2013'], 'answer_start': [3]}},

'question': "One of Beyonce's most successful tours yet was which one?",
'model_answer': '"The Formation World Tour"',
'gt_answer': {'text': ['The Mrs. Carter Show'], 'answer_start': [20]}},

'question': 'Where was Blue Ivy born?',
'model_answer': 'New York City',
'gt_answer': {'text': ['Lenox Hill Hospital in New York.'],

'question': 'When was 4 released?',
'model_answer': 'June 24, 2011',
'gt_answer': {'text': ['June 28, 2011'], 'answer_start': [42]}},

'question': 'Beyonce became the first female artist to perform solo in 20 years at which stage?',
'model_answer': 'Coachella',
'gt_answer': {'text': ['the 2011 Glastonbury Festival'],

'question': 'Which year did Beyonce and her father part business ways?',
'model_answer': '2011',
'gt_answer': {'text': ['2010'], 'answer_start': [60]}},

In few shot the model learnt to answer briefly which gave us 72 out of 200 exact matches, but the wording is sometimes different and sometimes the model answers incorrectly.

Generated knowledge prompt:

'question': 'When did Beyonce take a hiatus in her career and take control of her management?',
'model_answer': 'Beyoncé took a hiatus from music in 2010 and took over management of her career.',
'gt_answer': {'text': ['2010'], 'answer_start': [586]}},

Tanisha Khurana
Student ID - 200483139

'question': "What is the name of Beyoncé's younger sister?",
'model_answer': "Beyoncé's younger sister is named Solange.",
'gt_answer': {'text': ['Solange'], 'answer_start': [255]}},
{'id': '56d440df2ccc5a1400d830d4',
'question': 'Beyoncé is a descendant of which Acadian leader?',
'model_answer': 'Beyoncé is a descendant of Acadian leader Joseph Broussard.',
'gt_answer': {'text': ['Joseph Broussard.'], 'answer_start': [540]}},
{'id': '56d440df2ccc5a1400d830d5',
'question': 'Beyoncé was raised in what religion?',
'model_answer': 'Beyoncé was raised in a Methodist household.',
'gt_answer': {'text': ['Methodist'], 'answer_start': [578]}},

'question': "After leaving Destiny's Child, how many records did Beyoncé release under her own name?",
'model_answer': 'After leaving Destiny\'s Child, Beyoncé has released a total of six studio albums under her own name. These albums are:\n\n1. "Dangerously',
'gt_answer': {'text': ['118 million'], 'answer_start': [393]}},

In generated knowledge prompting, the model answered correctly but it was way too wordy.

The best method is combining few shot and generated knowledge prompting. To get the exact answers we need.