Question Generation with GPT-3 and Llama 2

Nitya Sampath
sampathnitya@gmail.com

Abstract

The task of question generation is to generate a question based on a given context and answer. It is useful for developing training data sets for QA models, among other uses. Recent developments in LLMs have shown impressive results in natural language generation, making them promising models for the task of question generation. Here, I explore the use of two LLMs, GPT-3 and Llama 2 on the task of question generation and investigate the effects of zero-shot and few-shot prompting on model performance. My experiments show that, despite expectations to the contrary, GPT-3 outperforms Llama 2 in both zero-shot and few-shot settings and that the performance of Llama 2 suffers when using few-shot prompting as opposed to zero-shot.

1 Introduction

Question generation is a natural language task where a model is provided a context passage and an answer and trained to generate a question that asks for the answer. The task is most commonly used to improve question-answering (QA) systems by increasing the size of the training data for the model. Other use cases include providing suggested or recommended questions to users in a question answering system, improving the conversational skills of chatbots by allowing them to generate leading questions, and even for educational purposes by automatically generating questions from reading material for assessment.

Early work on the area of question generation had focused on sequence to sequence models using such neural models as LSTMs, but with the advent of transformer models and LLMs, more recent techniques have utilized such models for this task given their versatility and their ability to understand and generate text over a wide range of tasks. GPT-3 (Brown et al., 2020) in particular has been a popular model for this research, as it has shown high performances on natural language generation tasks such summarization and question answering. Llama 2 (Touvron et al., 2023) is another widely used LLM that has shown comparable performance to GPT-3 on a number of language generation tasks including summarization and question answering. Given these results, it stands to reason that Llama 2 has the potential to match or even outperform GPT-3 on other, similar NLP tasks such as question generation.

Fine-tuning LLMs to perform on different tasks requires various prompt engineering techniques. In zero-shot learning, a model is prompted to perform without any task-specific examples. It relies solely on the instruction given in the prompt and its pre-existing knowledge to understand what is required and respond. Few-shot prompting, on the other hand, provides the model with a small amount of task-specific training data, called "shots". These shots are short examples that demonstrate to the model how to respond to the prompt. Using these demonstrations, the model is able to generalize to respond to new, unseen examples. Few-shot learning has been shown to improve the performance of LLMs because it allows the model to generalize from training examples.

This paper aims to build on previous work and investigate the use of LLMs (specifically, GPT-3 and Llama 2) and few-shot learning on the task of question generation. My core hypothesis for this paper is as follows:

Llama 2 will outperform GPT-3 on the task of question generation with both zero-shot and few-shot prompting.

I present a comparative analysis the two LLMs, GPT-3 and Llama 2 on the task of question generation. I prompted both models, using both zero-shot and few-shot techniques to generate questions based off of a given context and answer.

I analyzed their performances on context/answer pairs taken from the dataset SQuAD, using BLEU for quantitative analysis and conducted a qualitative analysis of the models' responses.

Results show that, in fact, Llama 2 does not match the performance of GPT-3 on this task. Furthermore, Llama 2's performance only declines with the use of few-shot prompting, contrary to the hypothesis.

2 Prior Literature

A common, early, approach to this problem of question generation has been a sequence to sequence approach utilizing neural models, specifically recurrent neural networks (RNNs) such as long short-term memory (LSTMs) or gated recurrent units.

Song et al. (2018), for instance, used a sequence-to-sequence model, with a birdirectional long short-term memory (BiLSTM) encoder to encode the passage and and an attentional LSTM decoder to generate the question. The encoder includes information about the answer position, but no other contextual information is given. Their architecture was motivated by and was intended to build on prior work (Sutskever et al., 2014), by making use of contextual information that previous models ignored. The authors trained their model on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). The authors found that their model outperformed both of the state of the art models they tested it against.

Zhou et al. (2019) sought to improve state of the art results on this task by building a model that could more accurately predict question-type, allowing it to then better generate a question appropriate to the target answer. They were motivated by previous work using models that would generate questions of the wrong type, as they could not properly predict what question type was most relevant to the target answer (Hu et al., 2018). The authors used a feature-rich encoder to encode the context and the answer, using POS tags, NER tags and word case as features. They passed the embeddings generated from this to a BiLSTM. For the question type prediction step, they classified questions into 8 types and used a unidirectional LSTM layer to predict the question types. Then, they used an attention based decoder to generate the question. Their model was tested on SQuAD and achieved state-of-the-art results.

More recently, BERT (Devlin et al., 2019) and various derived models have been used for this task, finetuning BERT models on question-answering datasets and using them to generate questions.

Chan and Fan (2019) were motivated to use BERT by building off of previous works which primarily used LSTMs (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Chung et al., 2014). The authors note that because of the limitations on these recurrent neural network (RNN) models processing long sequences, most of the previous models use sentence-level information for generating questions and show worse performance on paragraph length sequences. The state of the art model that the authors cite used maxout pointer and gated self attention mechanism, which enables it to process paragraphs (Zhao et al., 2018). The transformer architecture underlying BERT models allow for even more accurate processing of longer, paragraph level sequences. The authors tested their BERT-based models on SQuAD and achieved state of the art results.

Yin et al. (2021) also uses a BERT based model to tackle the task of question generation. Specifically, they used a BERT-based Pointer Generator Network (BERTPGN) as the decoder to generate questions. To encode the context and answer, they used positional and type embeddings as the input for the BERT model. Their aim was to generate questions that to recommended to used interacting with a QA system. Because of this use case, the authors sought to generate questions that were *self-explanatory* and *introductory*, which do not require background knowledge to understand the generated question and that are answered by longer passages rather than just simple statements. Their models were tested on the Natural Questions dataset (NQ) (Kwiatkowski et al., 2019), and ultimately improved on the results achieved by their gold standard reference.

Even more recently, and most related to the work presented in this paper, with the emergence of large language models, GPT-3 has been used to accomplish this task by prompting the model with the context passage and the answer and asking it to generate a question for the answer. Yuan et al. (2023) experimented with using GPT-3 and zero-shot prompting to generate questions based on context-answer pairs. They performed both a reference-based evaluation and a human evaluation of the results, acknowledging the limitations of a reference-based evaluation of generative models.

```
Given the context and an answer, provide a single question that can be answered by
the answer based on the context.

---

Follow the following format.

Context: will contain answer
Answer: ${answer}
Question: short question

---

Context: Despite waiving longtime running back DeAngelo Williams and losing top wide
receiver Kelvin Benjamin to a torn ACL in the preseason, the Carolina Panthers had
their best regular season in franchise history, becoming the seventh team to win at
least 15 regular season games since the league expanded to a 16-game schedule in
1978. Carolina started the season 14-0, not only setting franchise records for the
best start and the longest single-season winning streak, but also posting the best
start to a season by an NFC team in NFL history, breaking the 13-0 record previously
shared with the 2009 New Orleans Saints and the 2011 Green Bay Packers. With their
NFC-best 15-1 regular season record, the Panthers clinched home-field advantage
throughout the NFC playoffs for the first time in franchise history. Ten players were
selected to the Pro Bowl (the most in franchise history) along with eight All-Pro
selections.
Answer: Carolina Panthers
Question:  MODEL FILLS IN QUESTION HERE
```

Figure 1: Example zero-shot prompt, text before yellow was given as prompt

They found that in their reference-based evaluation, GPT-3 outperformed most of the other comparison models and that its responses were also scored highly by human evaluators, even when the model;s response was not highly similar to a reference or 'ground-truth' question.

This paper aims to build on this previous research, specifically the newer work with LLMs and test two of the top performing LLMs. My work seeks to provide a conparison of these widely used models on the task of question generation, and extending the analysis, not only to zero-shot, but also to few-shot prompting, to test the models on different styles of prompts.

3 Data

Like most of the prior work in this area, I used the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) for this paper. SQuAD is a commonly used dataset for question answering models and consists of over 100,000 question-answer pairs generated by crowdworkers on Wikipedia articles. The examples in the SQuAD dataset consist of the context passage, the question asked about the context and the answer to the question. The answer is a span from the corresponding context. SQuAD was originally a question-answering dataset, however, for this paper, I used it for question generation by providing the model with the context and the answer for each example. The original question for each example was kept aside to be used as a gold standard question for evaluation.

4 Model

I used GPT-3, prompted in a zero-shot manner, as a baseline. The primary models for comparison were then GPT-3 with few-shot prompting, and a Llama 2 model with both few-shot and zero-shot prompting. Both models were called through APIs. For GPT-3, I used GPT-3.5 Turbo, as it was the most cost effective GPT model that could provide results at a reasonable speed. For LLama2, I used Llama2-70b-Chat, which is a Llama 2 model fine-tuned for dialogue use cases.

5 Methods

5.1 Prompting

I used Stanford's DSPY framework (Khattab et al., 2023) to set up the models and the prompts. This required writing a signature for question generation and basic module to call the LLM in zero-shot setting. For few-shot prompting, I used a DSPY teleprompter to compile the basic module into a few-shot module using SQuAD's training split as data from which to draw the demonstrations.

I prompted the models to generate a question based on the given context and answer. The basic prompt was kept short, direct and succinct, in order to get the models to return an appropriate question without other, extraneous material. The same prompt was used for both GPT-3 and Llama2. The zero-shot prompts simply instructed the model to provide a single question, and included the context and answer the question was to be based on. Figure 1 shows an example of the basic zero-shot prompt, used for both models. For the few shot prompts, each prompt included 3 demonstrations, which were selected from the SQuAD training set, added into the basic prompt used for the zero shot experiments.

The models were tested on a sample of 400 examples randomly selected from the SQuAD dev set.

5.2 Evaluation

The responses for all four experiments (GPT3-zero-shot, GPT3-few-shot, Llama2-zero-shot, Llama2-few-shot) were evaluated by the BLEU metric, using the SQuAD example's original question as the gold standard for reference. Sentence BLEU was used to calculate the scores for each example for each model individually; corpus BLEU was used to calculate the BLEU score for each model over all 400 examples. The BLEU score represents the similarity between the reference sentence(s) and the example sentence, with a score of 1 meaning the example sentence is identical to the reference and a score of 0 meaning the example is completely different from the reference sentence.

Reference-based scoring of a task such as this is difficult because the models can return a question that is semantically similar in meaning, but still lexically different from the reference sentence. In addition, the questions and answers in the dataset were human generated, which means that the quality of the reference questions themselves might vary for each context. Because of this, I also completed a simple human evaluation of the the results. Each generated question was judged simply as "good" or "not good" based on whether it was an appropriate question to ask based on the answer and the context. Questions that were identical to the gold question were marked as "perfect".

6 Results and Analysis

Model	BLEU
GPT-3 (zero-shot)	0.4237
GPT-3 (few-shot)	0.4752
Llama 2 (zero-shot)	0.3355
Llama 2 (few-shot)	0.1766

Table 1: corpus BLEU scores for each experiment

Table 1 shows the overall corpus BLEU scores for each experiment. GPT-3 showed a better performance than Llama 2. While GPT-3 showed improvement with few-shot prompting, Llama 2 did not. These findings contradict the hypothesis that Llama 2 would outperform GPT-3 in both zero-shot and few-shot settings.

However, my human evaluation of the results showed that the gold questions and the questions returned by the models were different enough that the BLEU score did not always give a full picture of the quality of the individual returned questions. In many cases the models returned a question that was entirely appropriate for the example context and answer, but was worded differently or focused on a different part of the context than the gold question was, and therefore, received lower BLEU score.

6.1 GPT-3

6.1.1 Zero-shot (baseline)

Perfect	Good	Not Good	Total
6	353	41	400

Table 2: Human judgements for GPT-3 zero-shot

In the zero-shot setting, GPT-3 achieved a BLEU score of 0.4237, the second highest score of the 4 experiments. For the human evaluation judgements, Table 2 shows the number of generated questions that were judged as "good" and as "not good". GPT-3 generated 6 perfect questions with zero-shot, meaning that it returned a question identical to the gold question for 6 out of the 400 examples.

I grouped the results broadly into three categories: those that received a high BLEU score and were marked as "good", those that received a low BLEU score but were marked as "good", and those that received a low BLEU score and were marked as "not good".

Those in the first category, with a high BLEU score and a "good" judgement, were straightforward good responses by the model. These tended to be examples in which the answer was clearly defined by the context. The model would return the definition as the question. The "perfect" responses, those that received a BLEU score of 1 were all in this category and were all good questions for the answer based on the context.

The second category consisted of those examples for which the model received a low BLEU score, but the questions were judged as "good". These were responses that accomplished the task as expected, but did so differently from the gold answer. The differences between the GPT-3 generated responses and the gold questions can be grouped into two broad cases: longer questions and questions based off of a different part of the context from the gold question.

GPT-3 would consistently give longer questions, including extra information, than the gold questions. The gold questions, having been written by a human, tended to be shorter and simpler.

For example:

```
Context: Walt Disney and his brother Roy contacted Goldenson
at the end of 1953 for ABC to agree to finance part of the
Disneyland project in exchange for producing a television
program for the network. Walt wanted ABC to invest $500,000
and accrued a guarantee of $4.5 million in additional loans,
a third of the budget intended for the park. Around 1954, ABC
agreed to finance Disneyland in exchange for the right to
broadcast a new Sunday night program, Disneyland, which
debuted on the network on October 27, 1954 as the first of
many anthology television programs that Disney would
broadcast over the course of the next 50 years.
Answer: $500,000
Gold question: How much did Walt Disney want ABC to invest in
Disneyland?
GPT-3 question: How much did Walt Disney want ABC to invest
in the Disneyland project in exchange for producing a
television program for the network?
```

Figure 2: Example context & answer

Here, the model and the gold question were semantically very similar, and the primary difference was that the model included additional information that the gold question omitted. The generated question was grammatically correct and was answerable by the intended answer, but received a relatively low BLUE score of 0.4.

The second case of "good" questions with low BLEU scores consisted of examples in which multiple questions based off different facts in the context were possible and equally valid, and the model simply highlighted a different fact than the gold question did.

For example:

```
Context: Six-time Grammy winner and Academy Award nominee
Lady Gaga performed the national anthem, while Academy Award
winner Marlee Matlin provided American Sign Language (ASL)
translation.
Answer: the national anthem
Gold question: What did Marlee Matlin translate?
GPT-3 question: What did Lady Gaga perform at the event?
```

Figure 3: Example context & answer

Here, the model focused on a different part of the context than the gold question did, but still produced a question that was valid for the example. This response received a relatively low BLEU score of 0.3.

The third category were those that received a low BLEU score and were marked as "not good". These were entirely questions that were ungrammatical in form or not relevant to the context or answer. Most commonly, the model appears to misunderstand the context and was not able to generate a coherent question. This category, however, also included cases in which the model produced questions that were good questions solely for the given answer, but were not actually based off of the context, meaning that the model used its own knowledge to generate a question instead of relying on the context. These responses received a low BLEU score because they were very different from the gold answer from the dataset and I marked these as "not good" because the prompt asked the model to base the question off the context, which the model did not do.

6.1.2 Few-shot

Perfect	Good	Not Good	Total
8	349	43	400

Table 3: Human judgements for GPT-3 few-shot

With its score of 0.4752, GPT-3 in the few shot setting achieved the highest BLEU score out of the 4 experiments. GPT-3 few shot outperformed GPT-3 zero shot, which shows that few shot prompting did improve the results for this model, at least in BLEU scores.

The human judgement showed similar numbers for this few-shot experiment as for the zero-shot experiment. Accordingly, I found that the responses returned by GPT-3 in both zero-shot and few-shot were very similar. The same approaches to asking questions were followed, and in cases where the generated question differed from the gold questions, they typically differed in very similar ways.

For example:

```
Context: Six-time Grammy winner and Academy Award nominee
Lady Gaga performed the national anthem, while Academy Award
winner Marlee Matlin provided American Sign Language (ASL)
translation.
Answer: the national anthem
Gold question: What did Marlee Matlin translate?
GPT-3 zero-shot question: What did Lady Gaga perform at the
event?
GPT-3 few-shot question: What did Lady Gaga perform at the
event?
```

Figure 4: GPT 3 zero-shot vs. few-shot

Here, the responses from GPT 3 were exactly the same in zero-shot and few-shot and differed from the gold question in the same way.

Where the few-shot results did differe from the zero-shot, I found that the responses from GPT-3 few shot were generally clearer and better-formed questions, as compared to GPT-3 zero shot. They were more relevant to the context and the answer.

For instance:

```
Context: Gasquet (1908) claimed that the Latin name atra mors
(Black Death) for the 14th-century epidemic first appeared in
modern times in 1631 in a book on Danish history by J.I.
Pontanus: "Vulgo & ab effectu atram mortem vocatibant.
("Commonly and from its effects, they called it the black
death"). The name spread through Scandinavia and then Germany,
gradually becoming attached to the mid 14th-century epidemic as
a proper name. In England, it was not until 1823 that the
medieval epidemic was first called the Black Death.
Answer: atra mors
Gold question: What is the Latin name for Black Death?
GPT-3 zero-shot question: What Latin name did Gasquet claim was
first used for the 14th-century epidemic in modern times?
GPT-3 few-shot question: What is the Latin name for the Black
Death?
```

Figure 5: GPT 3 zero-shot vs. few-shot

The figure shows the context and answer for one of the examples in which GPT-3 showed improvement in the few-shot setting over the zero-shot setting.

Overall, I found that while few-shot and zero-shot resulted in very similar results, the "not good" few shot results were inappropriate in very similar ways as the "not good" zero-shot questions, while the "good" few-shot results were very often better than the "good" zero-shot questions. They were better constructed linguistically, clearer and more concise questions than those given by the zero-shot. For this reason, I conclude that few-shot prompting did improve the performance of GPT-3.

6.2 Llama 2

6.2.1 Zero-shot

Perfect	Good	Not Good	Total
2	343	55	400

Table 4: Human judgements for Llama 2 zero-shot

Llama 2 zero-shot achieved a corpus BLEU score of 0.3355. It did not outperform GPT-3, contradicting the hypothesis that Llama 2 would have a better performance. The human judgements of Llama 2's results also showed that it did not outperform GPT-3. Llama 2 had a higher number of "not good" responses and a lower number of "good" responses. It struggled with examples that GPT-3 was able to generate good questions for.

For instance, Figure 6 shows an examples that compares the results of Llama 2 zero-shot and GPT-3 zero-shot and few shot.

Here, GPT-3 was able to generate an appropriate question in both zero-shot and few-shot. However,

```
Context: Over the next five days, private conferences were
held to determine Luther's fate. The Emperor presented the
final draft of the Edict of Worms on 25 May 1521, declaring
Luther an outlaw, banning his literature, and requiring his
arrest: "We want him to be apprehended and punished as a
notorious heretic." It also made it a crime for anyone in
Germany to give Luther food or shelter. It permitted anyone to
kill Luther without legal consequence.
Answer: kill Luther
Gold question: What could anyone do to Martin Luther without
legal consequence?
GPT-3 zero-shot question: What was permitted by the final draft
of the Edict of Worms regarding Martin Luther?
GPT-3 few-shot question: What was anyone in Germany permitted
to do to Luther without legal consequence?
Llama 2 zero-shot question: What was the penalty for anyone
who killed Luther, according to the Edict of Worms?
```

Figure 6: GPT 3 vs Llama 2

Llama 2's question is not appropriate for the example, as it cannot be answered by the answer.

In addition, there were a number of responses from Llama 2 that were formatted in undesirable ways, such as restating the context and/or the answer in the response rather than simply returning the question as the prompt instructed it to do. In several cases, the response from Llama 2 looked like this:

```
Question: Who invented the Woolf high-pressure compound engine
in 1805?

Answer: British engineer Arthur Woolf.
```

Figure 7: example Llama 2 response with incorrect format

This is not expected behavior from the model. The prompt asks the model to simply return the question, as "Who invented the Woolf high-pressure compound engine in 1805?". GPT-3 did not show this behavior in either zero-shot ro few-shot.

Because of the lower scores with both BLEU and the human judgements and the issues with response formatting, Llama 2 did not outperform GPT-3 on this task.

6.2.2 Few-shot

Perfect	Good	Not Good	Total
3	220	177	400

Table 5: Human judgements for Llama 2 few-shot

Out of the four experiments run, Llama 2 in the few-shot setting had the lowest BLEU scores.

The vast majority of Llama 2's "not good" questions followed a similar format. The responses contained more information than the prompt was asking for.

```
Sure, here are the questions based on the given context and
answers:

1. What group did Paul VI address in New York in 1965?
2. What did Sander's study show in terms of black law students'
rankings?
3. What problems does linguistic anthropology bring linguistic
methods to bear on?
4. What is the name of the lake where the mouth of the Rhine
forms an inland delta?
```

Figure 8: example Llama 2 few-shot response

Figure 8 shows the typical "not good" answer from Llama 2 with few-shot. These responses include 4 questions rather than just the one asked for. The first three questions are the questions given as few-shot demonstrations, and only the last question is the generated question the prompt asked for. This was not behavior that was expected from the model. In my experiments, I attempted to adjust the prompt in order to avoid this behavior, but no matter the prompt, Llama 2 behaved this way on a signification number of examples. This behavior was not observed to this degree with Llama 2 in the zero-shot setting. GPT-3 did not show behavior like this, in either the zero-shot setting or the few-shot setting. It seems that the inclusion of the few-shot demonstrations does not help the model and only confuses its response to the task.

Out of the 177 responses judged "not good", 134 of them followed the pattern of including such extraneous information. However, as noted above, the model did generate questions for the test context and answer given for each of the examples it answered this way. In the response shown above, the question it generated was actually good, despite the response being wrongly formatted.

Perfect	Good	Not Good	Total
7	93	34	134

Table 6: Human judgements for Llama 2 wrong format responses

The table above shows the numbers for a human judgement of the Llama 2 responses that were originally formatted incorrectly. Llama 2's questions were generally of high quality, but got very low BLEU scores because of the format of the response. In addition, the model did not follow the prompt, as the prompt instruction was to return only the question and not any other information.

Ultimately, Llama 2 few-shot did not match or outperform GPT-3. The model generated questions well, but could not response the way the prompt instructed to. Because of this, its performance was worse as compared to GPT-3. Llama 2 also had decreased performance with few-shot than with zero-shot, which further contradicts the hypothesis that few-shot prompting would improve the model.

7 Conclusion

Here, I have presented a comparison of the performances of GPT-3 and Llama 2 on the task of question generation. I found that Llama 2 does not outperform GPT-3 on this task, contrary to the original hypothesis that it would. While the quality of Llama 2's generated questions was comparable to GPT-3's, Llama 2 was not as capable of following the instruction of the prompt, especially in few-shot, which meant that it returned worse results than GPT-3.

There are many angles future work might take to build on this paper. This paper only looked at GPT-3.5 Turbo and Llama-2-70 Chat. Both of these models are widely used in the field, but there are other variants of these models that could also be investigated, in particular GPT-4 or any of the other Llama family models, such as Llama 3.

In addition, the poor performance of Llama 2 being due to response formatting issues, another area of future work might look into further constraining model output to resolve the issues in formatting.

Known Project Limitations

This paper relies on reference-based evaluation and human judgements as evaluation metrics. Reference-based scores are not ideal for evaluating generative LLMs, given the fact that there are many equally valid possibilities for how the LLM might convey the same information. A response might be just as good, or even better than a reference, and still receive a low evaluation score because it used vastly different phrasing. In addition, I chose to use a very simple judgement method to compare the results of the different experiments, which may not capture the full nuance of the models results. I acknowledge that these evaluation methods constitute limitations that should be investigated further.

Authorship Statement

I was the sole author on this paper, responsible for running the experiments and collecting the data, the analysis and interpretation of the results and the writing of the paper. I did not receive any outside help.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Wenpeng Hu, Bing Liu, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2018. Aspect-based question generation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Xusen Yin, Li Zhou, Kevin Small, and Jonathan May. 2021. Summary-oriented question generation for informational queries. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 81–97, Online. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained LLMs: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada. Association for Computational Linguistics.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6032–6037, Hong Kong, China. Association for Computational Linguistics.