

Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation

Xingdi Yuan^{✳*} Tong Wang^{✳*} Yen-Hsiang Wang^{◇*}
Emery Fine[✳] Rania Abdelghani[✳] Hélène Sauzéon[✳] Pierre-Yves Oudeyer[✳]

[✳]Microsoft Research, Montréal [◇]National Chung Hsing University [✳]INRIA

eric.yuan@microsoft.com

wangtong106@gmail.com

heliart@smail.nchu.edu.tw

Abstract

Large Language Models (LLMs) have in recent years demonstrated impressive prowess in natural language generation. A common practice to improve generation diversity is to sample multiple outputs from the model. However, partly due to the inaccessibility of LLMs, there lacks a simple and robust way of selecting the best output from these stochastic samples. As a case study framed in the context of question generation, we propose two prompt-based approaches, namely round-trip and prompt-based score, to selecting high-quality questions from a set of LLM-generated candidates. Our method works without the need to modify the underlying model, nor does it rely on human-annotated references — both of which are realistic constraints for real-world deployment of LLMs. With automatic as well as human evaluations, we empirically demonstrate that our approach can effectively select questions of higher qualities than greedy generation.¹

1 Introduction & Related Work

Large Language Models (LLMs) have recently gained tremendous popularity in the NLP community (Devlin et al., 2019; Liu et al., 2019; Bao et al., 2020; Brown et al., 2020). The ever-increasing size in both models and training data renders many traditional learning methods impractical/intractable. As a result, prompt-based learning has emerged as a new paradigm tailored specifically towards leveraging the power of LLMs (Radford et al., 2019; Petroni et al., 2019; Raffel et al., 2020; Brown et al., 2020; Schick and Schütze, 2021b; Gao et al., 2021; Liu et al., 2021). In the zero-shot setting (such as in this study), a data sample is first “verbalized” into an input prompt and a ground-truth response in natural language. The prompt is then issued to a pre-trained LLM to obtain a predicted

response, which is then compared to the ground-truth for evaluation. This new technique has been successfully applied to many applications including text classification (Yin et al., 2019; Schick and Schütze, 2021a), QA (Jiang et al., 2021), natural language generation (Li and Liang, 2021) and NLG evaluation (Yuan et al., 2021).

Despite the impressive results on popular NLP benchmarks, however, the back-end LLMs are usually pre-trained with general-domain data, leading to sub-optimal performance in new domains for prompt-based learning. There are two major challenges in successful applying general-purpose LLMs to specific domains. Firstly, aside from the many known issues of LLMs (Webson and Pavlick, 2021; Min et al., 2022; Zhao et al., 2021; Lampinen et al., 2022), their sheer size and/or accessibility (e.g., served via API over the internet) makes it prohibitively expensive and impractical for domain adaptation. These limitations have inspired a recent line of work known as prompt editing/tuning (Gao et al., 2021; Li and Liang, 2021; Madaan et al., 2022). Additionally, prompt-tuning often relies on the availability of ground-truth labels of the data, imposing much additional resource on the approach.

Given the ubiquity of these challenges, our study focuses on alleviating the constraints on both annotation availability and access to model parameters, making LLMs more accessible for deployment. We take a mainstream NLG task, namely question generation, as a case study (Du et al., 2017; Yuan et al., 2017; Du and Cardie, 2018; Pan et al., 2019; Liu et al., 2020; Pyatkin et al., 2021). In this task, a model is trained to generate a natural language question conditioned on a context and an answer, such that the generated question can be answered by the provided answer using the context as supporting evidence. Question generation is the corner stone for many NLP applications in education (Kurdi et al., 2020; Abdelghani et al., 2022), FAQ

^{*} Equal contribution.

¹We open-source all code and annotated data on [github](#).

generation (Mass et al., 2020), information seeking (Qi et al., 2020), etc. In an educational setting, for example, a question generation system can generate demonstrations that inspire students’ curiosity and thinking (teaching), or to help assess students’ proficiency on certain knowledge or skills (examining). These use cases would benefit greatly from reduced dependency on computing resources, data availability, and the expertise required for fine-tuning an LM.

To align with these real-world scenarios, our goal is to obtain better outputs from an inference-only LLM (i.e., as a “black-box”, which is relatively more accessible, e.g., through online APIs). In particular, given the common practice of sampling multiple outputs to improve generation diversity, we propose a method that aims at selecting the best candidate based on multiple aspects of question quality in a zero-shot manner — notably without model adaptation or human annotations. Our method can be seen as a post-hoc selection process within a larger NLG pipeline, and thus is orthogonal and applicable to zero-shot and in-context learning methods (Rubin et al., 2021; Lu et al., 2022; Liu et al., 2022).

2 Problem Setting

Notations: Formally, we consider a dataset of context-answer pairs (c, a) both as strings. The task of question generation is to generate a question q that can be answered by a using c as supporting evidence. We use an off-the-shelf pre-trained LLM-based question generator in a zero-shot setting (prompt construction detailed in Appendix A). To simulate the black-box generator scenario, we refrain from any form of model tuning. We do, however, assume access to a set of output sequences stochastically sampled from the question generator. We thus ground our study to this application scenario by sampling k questions $Q = \{q_i : i = 1, \dots, k\}$. For comparison as a baseline, we also denote q_g as the question generated with a greedy algorithm (i.e., generating the most probable token at each time step).

Our goal is to devise an algorithm S which selects the best candidate q_{i^*} that maximizes some evaluation metric $M : Q \mapsto \mathbb{R}$, i.e., $S(Q) = i^* = \arg \max_i M(q_i)$. We use M_s , $M_{\bar{s}}$, and $M_{\underline{s}}$ to denote the mean, min, and max of $\{M(q) : q \in Q\}$, resp., and M_g for the greedy output $M(q_g)$. Semantically, $M_{\underline{s}} \leq M_s \leq M_{\bar{s}}$ is tautologically true,

and a positive result on the design of S would translate to $M(q_{S(Q)})$ outperforming both M_s and M_g .

Datasets and model: We adopt two question generation datasets with distinctive characteristics, namely SQuAD (Rajpurkar et al., 2016) and Fairytale QA (Xu et al., 2022). SQuAD was originally proposed as an extractive QA dataset. It has been used as a *sentence-level* question generation task in the question generation literature (Du and Cardie, 2018; Yuan et al., 2017; Bao et al., 2020), i.e., a context c is a single sentence that contains the corresponding answer a as a sub-string. Fairytale QA has also been used for both question answering and question generation. It features *paragraph-level* question generation (with c being one or more paragraphs), and the answer a is not necessarily a sub-string of c . Since we do not perform any form of model/prompt tuning, we use the testing split of both datasets, which consist of 11,877 data points for SQuAD and 1,007 for Fairytale QA.

We prompt GPT-3 (Brown et al., 2020)² in a 0-shot manner for both question generation and selection (detailed in §3). We provide all prompts in Appendix A.

Evaluation Metrics : We use two metrics to evaluate the selected question $q' = M(q_{S(Q)})$:

- Reference-based evaluation: Following prior works, we use BLEU-4 for SQuAD (Du and Cardie, 2018; Bao et al., 2020) and ROUGE-L for Fairytale QA (Xu et al., 2022)³. These metrics compare q' against the reference question \hat{q} (a.k.a. the “ground-truth” question in the existing literature).
- Human evaluation: we solicit human annotations on a subset of the data. We postulate that an overall score given holistically to rate a question would be highly subjective and thus less inductive to annotator agreement. Accordingly, we decompose the quality of questions into seven dimensions⁴, and ask human annotators to rate a question on each dimension followed by an overall rating of the question. We collect three annotations from different annotators for each data points. We provide details of the human study in Appendix B.

²We use the `text-davinci-002` model.

³We adopt evaluation scripts provided with the datasets.

⁴Namely, grammatical correctness, offensiveness, clarity, relevance, importance, specificity, and answerability.

[context]
 Old Dragonbeard must have been a master swordsman standing midway between those of the first and of the second order. Molo, however, of whom this story tells, was a sword hero.
 At that time there lived a young man named Tsui, whose father was a high official and the friend of the prince. And the father once sent his son to visit his princely friend, who was ill. The son was young, handsome and gifted. He went to carry out his father's instructions. When he entered the prince's palace, there stood three beautiful slave girls, who piled rosy peaches into a golden bowl, poured sugar over them and presented them to him. After he had eaten he took his leave, and his princely host ordered one of the slave girls, Rose-Red by name, to escort him to the gate. As they went along the young man kept looking back at her. And she smiled at him and made signs with her fingers. First she would stretch out three fingers, then she would turn her hand around three times, and finally she would point to a little mirror which she wore on her breast. When they parted she whispered to him: "Do not forget me!"

[question]
 Who was Rose-Red?

[input1]
 Is the [question] related to the [context]? Why?

[output1]
 Yes, the question is related to the context. Rose-Red is one of the three beautiful slave girls who served the prince. The young man, Tsui, was attracted to her and she seemed to be attracted to him as well.

[input2]
 Based on the above response [output1], which one of the following [options] best described [question] wrt [input1]?

[options]
 1: They are not at all related; 2: They are remotely related; 3: They are somewhat related; 4: They are closely related.

[output2]
 3: They are somewhat related.

Figure 1: Template for prompting GPT-3 to rate a question’s relevance. GPT-3 output is highlighted in green.

3 Method

In this section we propose three question selection methods. As described in §2, each method is used to score k sampled questions in Q and selects the candidate with the highest score.

n -gram similarity: We use n -gram similarity between a question and its context to measure their relevance. This method reflects the intuitive assumption that favorable question be closely related to the information provided by the context. Specifically, we extract all unique n -grams⁵ $s^n(c)$ from a given context c , $s^n(q)$ from a question q . The n -gram similarity score is then defined as:

$$\text{sim}^n = \frac{|s^n(c) \cap s^n(q)|}{|s^n(q)|}, \quad (1)$$

where $|s|$ indicates the size of set s .

Round-trip: Intuitively, the answer to a generated question should be semantically equivalent to the answer that has been used to generated the question. Formally, a question generation model QG and a QA model (both with reasonable performance) should satisfy the following constraint:

$$q' = \text{QG}(c, a); \quad a' = \text{QA}(c, q'); \quad a' = a. \quad (2)$$

This idea is closely related to *cycle consistency* in the existing literature on image generation (Zhu et al., 2017), machine translation (Artetxe et al., 2018), and QA (Alberti et al., 2019; Shah et al., 2019)). Here, we use GPT-3 as an off-the-shelf QA model to obtain a' for each pair of c and q' , resulting in k answers $A = \{a'_1, \dots, a'_k\}$ for the k

⁵In all our experiments n ranges from 1 to 5.

	SQuAD (BLEU-4)	Fairytale QA (ROUGE-L)
prior works (models trained/fine-tuned on these datasets)		
(Du and Cardie, 2018)	0.152	–
(Zhang and Bansal, 2019)	0.184	–
UniLM Large (Bao et al., 2020)	0.228	–
UniLM v2 Base (Bao et al., 2020)	0.244	–
ERNIE-GEN Large (Xiao et al., 2021)	0.254	–
BART (Xu et al., 2022)	–	0.527
baselines (notations defined in §2)		
M_g (greedy)	0.372	0.424
M_s (sample avg)	0.359	0.399
$M_{\underline{s}}$ (lowerbound, from exhaustive search)	0.225	0.259
$M_{\overline{s}}$ (upperbound, from exhaustive search)	0.496	0.548
question selection		
bi-gram	0.382	0.403
tri-gram	0.380	0.403
round-trip	0.392	0.434
overall prompt score (OPS)	0.373	0.399
averaged prompt score (APS)	0.380	0.406
ensemble multiple methods		
APS + round-trip	0.397	0.439
bi-gram + round-trip	<u>0.400</u>	0.429
tri-gram + round-trip	0.398	0.430
bi-gram + APS	0.384	0.406
tri-gram + APS	0.383	0.409
bi-gram + APS + round-trip	0.401	0.431
tri-gram + APS + round-trip	<u>0.400</u>	<u>0.435</u>

Table 1: Reference-based evaluation scores. Best and second best scores (excluding baselines) are highlighted with **boldface** and underline.

sampled questions in Q . We then measure the similarity between each a'_i and the ground-truth answer a (F_1 for SQuAD and ROUGE-L for Fairytale QA, in accordance with the evaluation setup from the original papers for the two datasets). Finally, we select the question corresponding to the generated answer a'_i that overlaps the most with a (i.e., that can be best answered by GPT-3). Prompts used in these experiments are detailed in Appendix A.

Prompt-based Score: We propose a two-step procedure (Figure 1) for prompting GPT-3 to answer the same set of meta-questions (i.e., questions about the quality of a given question) used for human evaluation (§2). In step 1, given a context-question pair, GPT-3 is prompted to answer a meta-question as an open question (as opposed to choosing among a list of options) as well as to verbalize a reason for its answer. In step 2, GPT-3 is prompted to choose from a list of options representing the rating scale of the meta-question.

We empirically observe that without the first step, GPT-3 output tends to have a low-entropy distribution, i.e., often choosing the same option for a given meta-question disregarding the different context-question pairs. In contrast, the additional first-step appears to improve prediction diversity, which is inline with observations made in some existing studies (Nye et al., 2021; Wei et al., 2022).

Similar to human evaluation, we also prompt

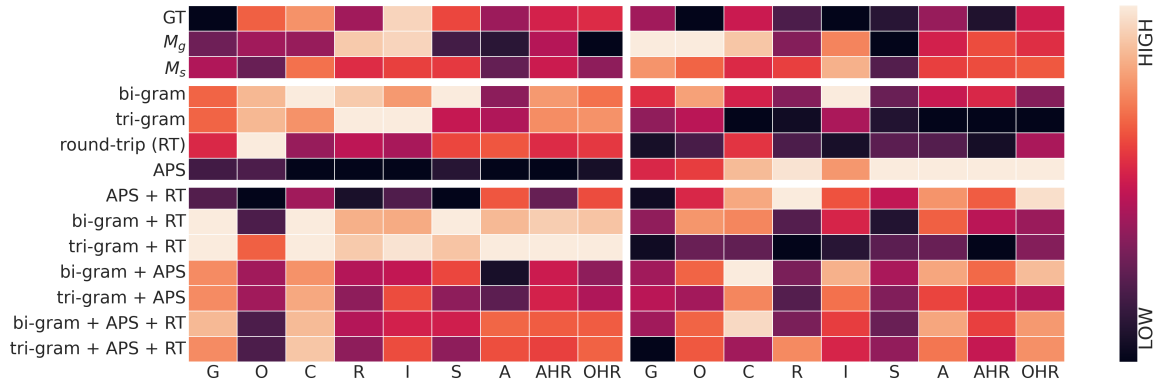


Figure 2: Human evaluation results, averaged over three annotators’ scores, normalized per column. Left: SQuAD; right: Fairytale QA. Abbreviations in x-axis denote **G**rammatical correctness, **O**ffensiveness, **C**larity, **R**elevance, **I**mportance, **S**pecificity, **A**nswerability, **A**veraged **H**uman **R**ating (over all dimensions to the left), **O**verall **H**uman **R**ating (an overall score given by annotators). Exact scores are provided in Appendix C.

GPT-3 to generate an overall score of a question (denoted *overall prompt-based score* or *OPS*). The average score of all individual meta-questions is instead denoted *averaged prompt-based score* (*APS*).

4 Results and Discussion

To measure the performance of a selection method (§3), we use it to select one out of $k = 5$ questions sampled from GPT-3, and score the selection with the evaluation metrics outlined in §2. Additionally, we test the ensemble performance with multiple methods. To ensure comparability, we normalize the scores obtained from each selection method into the range between 0 and 1, and use their average score to perform question selection.

Reference-based evaluation Reference-based evaluation are automatic metrics that are applied to the entire test sets of SQuAD and Fairytale QA. Table 1 shows that on both datasets, all question selection methods outperform M_s , the average score over all five sampled questions, validating the effectiveness of the proposed methods. While all individual methods outperform the greedy baseline M_g on SQuAD, round-trip performs the best, outperforming M_g on both datasets. It can be further improved via ensemble with n -gram and/or prompt-based scores (using uniform weights).

Note that prior studies require a large amount of labeled data for model training/fine-tuning, while GPT-3 performs zero-shot inference. Despite this major difference in learning paradigm, most GPT-3-based models proposed here outperform previous results by significant margins on the SQuAD dataset — even the least performant samples M_s (lowerbound) achieve competitive results. For

Fairytale QA, however, only the best samples M_s (upperbound) outperform previous results (Xu et al., 2022), indicating margins for improvement on question selection strategies for future work.

Human evaluation Human evaluation consists of 16, 800 annotations (from 87 annotators) evenly split across the two datasets (details in Appendix B). For question generation (among many NLG tasks), model outputs may exhibit linguistic diversity while maintaining semantic equivalence. It is thus highly problematic to evaluate such outputs against a single reference (i.e., “ground-truth”). Figure 2 empirically shows that the ground-truth (GT) provided in the datasets often fail to receive the highest human ratings, on many occasions scoring lower than stochastic samples from GPT-3 (M_s). Consequently, we strongly advocate for human evaluation, which we believe is highly effective in improving generalizability of our results to real-world applications.

Another prominent observation is that n -gram and APS perform quite differently on the two datasets. On SQuAD, n -gram similarity outperforms other individual methods, with further noticeable improvements via ensemble with round-trip. APS, on the other hand, does not work nearly as well, performing the worst for almost all meta-questions. In contrast, n -gram (particularly tri-gram) similarity shows the worst performance on Fairytale QA, while APS outperforms all other methods by a noticeable margin.

We posit that the reversed trend in comparing n -gram and APS can be explained by the distinct natures of the datasets. For SQuAD, the sentence-level contexts are relatively short and simple with

strictly extractive answers (i.e., the answers being sub-strings of the corresponding contexts). As a result, paraphrasing the context can be a fairly effective strategy to generate questions (hence the stronger correlation between question quality and the c - q n -gram similarity). In contrast, with multi-paragraph contexts and abstractive, open-ended answers, Fairytale QA questions are more likely posed about abstract ideas rather than simple context paraphrasing. Consequently, n -gram similarity, which favors local context paraphrasing, is less likely to perform well.

Conclusion In this study, we investigate the practical problem of selecting the best output from multiple samples generated by an LLM. Using question generation as a case study, we propose two prompt-based question selection methods. To alleviate real-world constraints on using LLMs, the proposed methods do not require model fine-tuning nor human annotation. Extensive experiments with both automatic and human evaluations evince the effectiveness of our approach on question selection.

5 Limitations

We acknowledge that our system has some limitations that warrants further investigation. For example, one needs to be mindful of the specific downstream applications of the proposed methods, both in terms of 1) potentially large variance in out-of-distribution performance (e.g. *divergent* question generation applications that aim to spark children’s curiosity-driven thinking (Abdelghani et al., 2022)); and 2) of mitigating harmful/toxic contents in educational applications (Bender et al., 2021). As a result, we believe such techniques and applications are neither suitable nor safe to directly interact with children, we urge developers to use this technique in other ways, for instance, in teaching assistant application (e.g., a system that suggests examples for teachers), where the teacher can filter and modify the examples and thus making sure the content children receive is proper and safe.

We also acknowledge the prohibitively restrictive access to the GPT-3 model at the time of writing. We do believe that this constraint will relax over time, and meanwhile, hoping that our proposal can shed light on research and applications with more accessible LLMs such as GPT-J (Wang and Komatsuzaki, 2021) and BLOOM (BigScience, 2022) for future work.

While we acknowledge the many limitations

with respect to accessing GPT-3, we are not advocating against using it. On the contrary, in fact, we believe GPT-3 is still among the most cost-effective solutions especially in the context of natural language generation. The main goal of the study is thus to explore more data efficient ways of using GPT-3 to generate and evaluate questions. We strive to share our experience and insights with the community, which hopefully can be proven valuable and helpful.

References

- Rania Abdelghani, Pierre-Yves Oudeyer, Edith Law, Catherine de Vulpillieres, and H elene Sauz eon. 2022. Conversational agents for fostering curiosity-driven learning in children. *International Journal of Human-Computer Studies*.
- Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. [Synthetic QA corpora generation with roundtrip consistency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. [Unsupervised neural machine translation](#). In *International Conference on Learning Representations*.
- Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. [UniLMv2: Pseudo-masked language models for unified language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- BigScience. 2022. Bigscience language open-science open-access multilingual (bloom) language model. *International*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- 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.
- Xinya Du and Claire Cardie. 2018. [Harvesting paragraph-level question-answer pairs from Wikipedia](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.
- Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. [Making pre-trained language models better few-shot learners](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3? In Proceedings of Deep Learning Inside Out \(DeeLIO 2022\): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures](#), pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*.
- Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised faq retrieval with question generation and bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812.
- Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.
- Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

- Valentina Pyatkin, Paul Roit, Julian Michael, Reut Tsarfay, Yoav Goldberg, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. *arXiv preprint arXiv:2109.04832*.
- Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. *Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations*. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of Machine Learning Research*, 21(140):1–67.
- 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.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.
- Timo Schick and Hinrich Schütze. 2021a. *Exploiting cloze-questions for few-shot text classification and natural language inference*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021b. *It’s not just size that matters: Small language models are also few-shot learners*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.
- Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.
- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa—an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. *Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach*. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. *BartScore: Evaluating generated text as text generation*. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.
- Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordani, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *arXiv*.
- Shiyue Zhang and Mohit Bansal. 2019. *Addressing semantic drift in question generation for semi-supervised question answering*. 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 2495–2509, Hong Kong, China. Association for Computational Linguistics.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Contents in Appendices:

- In Appendix A, we report all prompt templates we used in this work.
- In Appendix B, we provide details on the human study.
- In Appendix C, we provide the full set of our experiment results.
- In Appendix D, we report implementation details.

A Prompt Designs

We report an example of our prompt for question generation in Figure 3.

We report an example of our prompt for QA (used in round-trip) in Figure 4.

We report an example of our prompt in obtaining prompt scores in Figure 1.

B Human Study

We randomly sample 50 documents from each of the two datasets SQuAD and Fairytale QA. Each document correspond to one ground-truth question and six questions generated by GPT-3 (five by stochastic sampling and one by greedy search). Each question is then rated by three human annotators wrt seven meta-questions and one over-all rating, altogether constituting $50 \times 2 \times (1 + 5 + 1) \times 3 \times (7 + 1) = 16,800$ annotations. There are in total 87 annotators involved in the annotation process, all annotators are English speakers, they are recruited from regions including Europe, the United States and United Kingdom. Each annotator on average performed 193 annotations and was paid on average \$14.1 USD per hour.

We perform a basic spam filtering process on the raw annotations. We observe a 15.4% spam rate. All human scores reported in this paper are computed after spam removal.

We report the eight meta-questions we used for human annotation in Figure 5. The eight meta-questions correspond to columns in Figure 2. We collect three annotations from different annotators for every meta-question, we report the averaged human agreement rate in Table 2.

C Additional Results

In Table 3, we report the full experiment results for reference-based evaluation.

grammatical correctness	0.698
offensiveness	0.788
clarity	0.640
relevance	0.670
importance	0.558
specificity	0.619
answerability	0.588
overall human rating (OHR)	0.485

Table 2: Averaged human agreements among three annotators. An agreement indicates that all three annotators selected the same option for a meta-question. We show decomposing single-score metric (i.e., OHR) to scores measuring different aspects (listed above OHR) can significantly improve human agreements.

In Table 4, we report the full results for human evaluation on SQuAD.

In Table 5, we report the full results for human evaluation on Fairytale QA.

D Implementation Details

In all experiments, we use the `text-davinci-002` (175B parameters) variant of GPT-3. It is currently the most capable GPT-3 model variant. Compared to other variants, `text-davinci-002`'s support to inserting completions can better facilitate our question generation tasks (as shown in Figure 3).

We use a temperature of 0.7 during the sampling process of question generation. In all other use cases (e.g., QA round-trip, prompt score), we use greedy generation (temperature is set to 0).

Story:

As soon as the lady had departed the fisher's son awoke, and the dark lad told him of her visit, and how he would never see her as long as he lived. At this the fisher's son felt the cold creeping up to his heart, yet he knew the fault had not been his that sleep had overtaken him. 'I will search the whole world through till I find her,' cried he, and the dark lad laughed as he heard him. But the fisher's son took no heed, and off he went, following the sun day after day, till his shoes were in holes and his feet were sore from the journey. Nought did he see but the birds that made their nests in the trees, not so much as a goat or a rabbit. On and on and on he went, till suddenly he came upon a little house, with a woman standing outside it.

Instruction:

Read the above story, ask a question and answer it.

Question:

GPT-3 FILLS IN THIS BLANK

Answer:

search the whole world through till he found her

Figure 3: An example of prompting GPT-3 for question generation. We use the text before `green` as prompt, and text after `green` as suffix. We refer readers to [the GPT-3 documentation](#) for more details about GPT-3's inserting completion mode.

[Document]:

is cheeks were red with passion, and his eyes were bright, for he could not but notice that, now that she was safe at Orphir under her true love's protection, the Lady Morna's manner had grown cold and distant again, and he was beginning to lose faith in Snorro's charm.

Angry and disappointed, he had sought his mother's room to pour out his story of vexation to her.

He stopped short, however, when he saw the wonderful waistcoat lying on the table, all gold and silver and shining colours. It was like a fairy garment, and its beauty took his breath away.

[Question]:

Why did Harold lose faith in Snorro's charm?

[Answer]:

Harold lost faith in Snorro's charm because the Lady Morna's manner had grown cold and distant again.

Figure 4: An example of prompting GPT-3 for QA. GPT-3 output is highlighted in green.

1. Is the question grammatically correct?
 - 1) It is grammatically incorrect
 - 2) It has some grammatical issues
 - 3) It is grammatically correct

2. Is the question offensive to people?
 - 1) It is very offensive
 - 2) It may be offensive
 - 3) It is not at all offensive

3. Is the question clear?
 - 1) It is not at all clear
 - 2) It is mostly clear
 - 3) It is very clear

4. Is the question related to the context of the attached document?
 - 1) It is not at all related
 - 2) It is somewhat related
 - 3) It is closely related

5. Is the question asking about an important aspect of the context of the attached document?
 - 1) Not at all important
 - 2) It may be important
 - 3) It is very important

6. Is the question asking about a specific piece of information in the attached document?
 - 1) The question is very generic
 - 2) The question is somewhat generic
 - 3) The question is very specific

7. Can the question be answered using information in the attached document?
 - 1) No, answering the question requires completely different information
 - 2) The question can be partially answered using information from the document
 - 3) The question can be perfectly answered using information from the document

8. What is your overall rating of the question generated based on the attached document?
 - 1) The question is very bad
 - 2) The question is okay
 - 3) The question is very good

Figure 5: Meta-questions we designed for human evaluation.

	SQuAD (BLEU-4)	Fairytale QA (ROUGE-L)
prior works (models trained/fine-tuned on these datasets)		
(Du and Cardie, 2018)	0.152	–
(Zhang and Bansal, 2019)	0.184	–
UniLM Large(Bao et al., 2020)	0.228	–
UniLM v2 Base(Bao et al., 2020)	0.244	–
ERNIE-GEN Large (Xiao et al., 2021)	0.254	–
BART (Xu et al., 2022)	–	0.527
baselines (notations defined in §2)		
M_g (greedy)	0.372	0.424
M_s (sample avg)	0.359	0.399
M_s (lowerbound)	0.225	0.259
M_s (upperbound)	0.496	0.548
n -gram-similarity		
uni-gram w/ context	0.382	0.396
bi-gram w/ context	0.382	0.403
tri-gram w/ context	0.380	0.403
4-gram w/ context	0.378	0.406
5-gram w/ context	0.375	0.404
round-trip		
round-trip	0.392	0.434
prompt scores		
grammatical correctness	0.364	0.405
offensiveness	0.374	0.403
clarity	0.373	0.406
relevance	0.372	0.396
importance	0.372	0.406
specificity	0.378	0.405
answerability	0.372	0.404
averaged prompt score (APS)	0.380	0.406
overall prompt score (OPS)	0.373	0.399
ensemble multiple methods		
APS + round-trip	0.397	0.439
bi-gram + round-trip	<u>0.400</u>	0.429
tri-gram + round-trip	0.398	0.430
bi-gram + APS	0.384	0.406
tri-gram + APS	0.383	0.409
bi-gram + APS + round-trip	0.401	0.431
tri-gram + APS + round-trip	<u>0.400</u>	<u>0.435</u>

Table 3: Reference-based evaluation scores on various question selection methods. Best and second best numbers (excluding baselines) are highlighted with **boldface** and underline.

	G	O	C	R	I	S	A	AHR	OHR
GT	0.937	0.987	0.943	0.930	0.925	0.922	0.887	0.933	0.870
M_g	0.950	0.983	0.927	0.953	0.925	0.905	0.870	0.930	0.833
M_s	0.957	0.981	0.940	0.937	0.909	0.921	0.879	0.932	0.857
bi-gram	0.968	<u>0.990</u>	0.952	<u>0.953</u>	0.918	0.937	0.885	0.943	0.880
tri-gram	0.968	<u>0.990</u>	0.943	0.957	0.928	0.917	0.890	0.942	0.885
round-trip (RT)	0.962	0.992	0.927	0.933	0.900	0.922	0.903	0.934	0.872
APS	0.945	0.980	0.912	0.912	0.880	0.902	0.863	0.913	0.837
APS + RT	0.947	0.977	0.928	0.915	0.890	0.898	0.903	0.923	0.875
bi-gram + RT	0.983	0.980	0.952	0.950	0.920	0.937	<u>0.917</u>	<u>0.948</u>	<u>0.893</u>
tri-gram + RT	0.983	0.987	0.952	<u>0.953</u>	<u>0.927</u>	<u>0.933</u>	0.925	0.951	0.900
bi-gram + APS	0.972	0.983	0.943	0.932	0.903	0.922	0.867	0.932	0.857
tri-gram + APS	0.972	0.983	0.945	0.928	0.910	0.912	0.878	0.933	0.862
bi-gram + APS + RT	<u>0.977</u>	0.980	0.947	0.932	0.905	0.918	0.905	0.938	0.877
tri-gram + APS + RT	<u>0.972</u>	0.980	<u>0.948</u>	0.928	0.910	0.912	0.902	0.936	0.878

Table 4: Human eval results (SQuAD). Abbreviations in the first row denote **G**rammatical correctness, **O**ffensiveness, **C**larity, **R**elevance, **I**mportance, **S**pecificity, **A**nswerability, **A**veraged **H**uman **R**ating (over all dimensions to the left), **O**verall **H**uman **R**ating (an overall score given by annotators). Best and second best numbers (excluding baselines) are highlighted with **boldface** and underline.

	G	O	C	R	I	S	A	AHR	OHR
GT	0.945	0.963	0.942	0.937	0.885	0.928	0.892	0.927	0.867
M_g	0.975	1.000	0.958	0.943	0.920	0.922	0.905	0.946	0.870
M_s	0.964	0.988	0.944	0.955	0.925	0.934	0.912	0.946	0.875
bi-gram	0.953	0.993	0.943	0.943	0.932	0.937	0.902	0.943	0.857
tri-gram	0.943	0.980	0.922	0.930	0.905	0.927	0.858	0.924	0.838
round-trip (RT)	0.928	0.970	0.945	0.937	0.888	0.935	0.878	0.926	0.862
APS	<u>0.952</u>	0.985	0.957	<u>0.972</u>	0.922	0.977	0.948	0.959	0.895
APS + RT	0.927	0.983	0.955	0.973	0.915	<u>0.948</u>	0.928	0.947	<u>0.893</u>
bi-gram + RT	0.943	<u>0.992</u>	0.952	0.938	0.910	<u>0.927</u>	0.918	0.940	<u>0.860</u>
tri-gram + RT	0.927	<u>0.973</u>	0.932	0.928	0.890	0.935	0.882	0.924	0.857
bi-gram + APS	0.945	0.988	0.962	0.942	<u>0.925</u>	0.945	<u>0.932</u>	<u>0.948</u>	0.888
tri-gram + APS	0.948	0.978	0.952	0.938	<u>0.918</u>	0.940	<u>0.913</u>	<u>0.941</u>	0.863
bi-gram + APS + RT	0.945	0.988	<u>0.960</u>	0.942	0.913	0.937	<u>0.932</u>	0.945	0.883
tri-gram + APS + RT	0.925	0.987	0.938	0.962	0.910	0.942	0.922	0.941	0.882

Table 5: Human eval results (Fairytale QA). Abbreviations in the first row denote **G**rammatical correctness, **O**ffensiveness, **C**larity, **R**elevance, **I**mportance, **S**pecificity, **A**nswerability, **A**veraged **H**uman **R**ating (over all dimensions to the left), **O**verall **H**uman **R**ating (an overall score given by annotators). Best and second best numbers (excluding baselines) are highlighted with **boldface** and underline.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 5 on page 5.
- A2. Did you discuss any potential risks of your work?
Section 5 on page 5.
- A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 on page 1.
- A4. Have you used AI writing assistants when working on this paper?
This paper discuss an application of GPT-3, but we do not use GPT-3 in any of the paper section writing.

B Did you use or create scientific artifacts?

We use existing language model, namely GPT-3; we also use existing question generation dataset, namely SQuAD and Fairytale QA. We discuss them in Section 2 (page 2).

- B1. Did you cite the creators of artifacts you used?
Yes, we cite the creators in Section 2, page 2.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. They are publicly available.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in Section 1 and 5.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in section 2, 3 and appendix B.

C Did you run computational experiments?

Yes, described in section 2, 3, 4.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. We do not propose new model. We use existing language model, and properly cite the original work.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Not applicable. Left blank.

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Not applicable. Left blank.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Not applicable. Left blank.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, section 4 and appendix B.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Yes, appendix B.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Yes, appendix B.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Yes, appendix B.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Yes, appendix B.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Not applicable. Left blank.