Towards Reliable LLM-based Exam Generation

Lessons Learned and Open Challenges in an Industrial Project

Renzo Degiovanni
Luxembourg Institute of Science and Technology (LIST)
renzo.degiovanni@list.lu

Jordi Cabot
Luxembourg Institute of Science and Technology (LIST)
jordi.cabot@list.lu

***Abstract*—Large Language Models (LLMs) have revolutionized the way natural language tasks are handled, with big potential applications in the context of education. LLMs can save educators time and effort, for instance, in content creation and exam generation. Although promising, LLMs' integration into educational products brings some risks that companies must mitigate.**

In the context of an industrial project, we investigate the effectiveness of LLMs to generate educational multiple-choice questions. The experiments include 16 commercial and open-source LLMs, rely on standard metrics to assess the accuracy (F1 and BLEU) and linguistic quality (perplexity and diversity) of the generated questions, and compare with five specialized models. The results suggest that recent LLMs can outperform the fine-tuned models for question generation, open-source LLMs are very competitive with the commercial ones, with Meta Llama models being the best performing, and DeepSeek as performing as recent GPT4 models.

This promising empirical evidence encourages us to focus on advanced prompting strategies, for which we report relevant open challenges we aim to address in the short term.

I. Introduction

Recent advances in AI brought by the release of powerful Large Language Models (LLMs) have revolutionized the way natural language tasks are handled. LLMs can produce human-like text as well as images, audio and videos, and their novelty and creativity can provide valuable assistance [24], [9]. While they can be game-changer in several activities, concerns about their impact on education have been raised [3]. Among the potential benefits, LLMs can support educators to save valuable time and effort in creating and grading student exams. Although promising, LLMs are prone to hallucinations [10] and biases [7], which constitutes a big risk for companies aiming to integrate LLMs into their educational products.

In this paper, we study the performance of LLMs for *Educational Question Generation* [5], the problem of automatically generating questions from a given textual content and the desired response. This is done in the context of an industrial project [1] with the company Open Assessment Technologies S.A. (OAT[1]) that provides advanced assessment solutions for education, with users in 194 countries, in more than 30 languages, leading to millions of tests worldwide. The key objective of this project is to develop reliable and trustworthy LLM-based applications to assist teachers in the generation of educational exams. The application to be developed should provide assurances on the quality of the generated items and thus, we need to guide LLMs to produce items that are correct, free of biases, and calibrated for the target audience. Similarly to other applications in related domains, we seek to use LLMs for both, generating the exams and judging the quality of the generated content (i.e. LLMs-as-a-Judge [8]). However, there is not much evidence of recent LLMs performance for educational question generation. Hence, to identify concrete challenges we need to cope in the short-term, we design, as the initial step in our project, an empirical study to understand the basic capabilities of LLMs to generate educational questions.

Our experiments include 16 commercial and open-source LLMs from GPT, Meta Llama, Mistral AI, and DeepSeek. Our empirical evaluation is based on the standard SciQ [26] dataset, and the Canterbury Question Bank (CQB) [18] to study the generalization of the results. We rely on standard metrics for assessing question generation models, and measure the accuracy (F1 and BLEU scores) and the linguistic quality (perplexity and diversity) of the LLM generated questions. We use as a baseline five fine-tuned models, specialized in educational question generation, namely, Leaf [22], EduQG [5], and their variants Leaf+ and EduQG+, and RLLM-EduQG [11], that constitutes the current state-of-the-art.

The empirical results show that Meta Llama models (e.g. Llama-3.1-70B-Instruct) are the best performing models and can outperform all the fine-tuned models used as a baseline. Moreover, GPT-3.5 Turbo and GPT-4o can perform as well as the state-of-the-art model RLLM-EduQG [11], but other GPT versions (e.g. GPT-4o-mini) behave slightly worse than RLLM-EduQG, outperforming anyway the rest of the baselines. Mistral and DeepSeek models have a similar accuracy as GPT-4o-mini, and questions generated with DeepSeek are the ones with the best linguistic quality scores.

Our empirical results are promising and suggest that more elaborated prompt engineering techniques, such as Chain-of-Thought [17] and RAG [12], can improve LLMs' effectiveness even further. However, as discussed with our industrial partner, the textual context might not be sufficient to determine the adequacy and quality of the generated questions. Specific contextual information of the target audience, such as demographic information of the students and the expected difficulty, should be taken into account, and the questions should be adapted accordingly. Thus, we have identified relevant open challenges that we seek to address in the short term.

[1] https://www.taotesting.com/

2643-1572/25/$31.00 ©2025 IEEE
DOI 10.1109/ASE63991.2025.00313

On the one hand, in collaboration with the company's domain experts, we are analyzing the relevant dimensions we should take into account when designing our prompting strategies. We are in the process of defining a taxonomy that captures the target students' meta-data, to improve LLMs guidance towards student's profile-specific questions.

On the other hand, standardized exams have associated a diverse set of quality criteria that are relevant for our industrial partner. For instance, in multiple-choice questions, if the distractors are very similar to the correct answer, the question becomes ambiguous and unsuitable for assessment activities. We are developing specific *validators* that rely on LLMs-as-a-Judge strategies to identify issues in the generated questions that do not comply with the expected quality properties. We also plan to combine these validators with advanced testing techniques, such as metamorphic testing [19], to increase their robustness and reliability. Metamorphic transformations can be used to inject issues in the questions (a.k.a. mutants) that should be identified by the validators (e.g., by selecting a distractor as the correct answer), while other kinds of transformations, such as replacing a word by a synonym, should not change the semantic of the question.

In the remaining, we present our empirical study results, and discuss the open challenges and roadmap of our project.

II. Background and Related Work

The Educational Question Generation (EduQG) problem consists of automatically generating the question from a given support text. There are variants that also provide the expected correct answer as an input to narrow down the generation process [6]. Recent approaches for EduQG have focused on fine-tuning and leveraging pre-trained language models to improve their performance. We use the following customized models as baselines when assessing LLMs' effectiveness.

Leaf [22], a recent question generation system, fine-tunes a pre-trained language model (Google T5 [15]) on the SQuAD 1.1 dataset [16] for the question and multiple-choice distractor generation. EduQG [5] additionally pre-trains the model with scientific text documents, taken from the S2ORC dataset [13], before fine-tuning it for question generation. Leaf+ and EduQG+ extend these models by further fine-tuning the models with an educational question dataset, SciQ [26], specialised in general-purpose questions. RLLM-EduQG [11] fine-tunes the Google FLAN T5 model, on the SciQ dataset, using a mixed objective function that uses Reinforcement Learning to improve the generation of questions that are syntactically and semantically accurate.

Though effective, these models are not publicly available which limits their application in practice. However, nowadays we can easily interact with powerful LLMs through their intuitive interfaces (chatbots) and APIs. Thus, without requiring any fine-tuning, we can instruct the LLMs, via simple prompts, for their assistance in generating educational questions. In the following section, we study whether such general LLMs have the ability to provide results comparable to the quality of the questions generated by specialized models.

III. LLMs Effectiveness for Educational Question Generation

We start our analysis by evaluating the capabilities of LLMs for educational question generation. Hence, we ask:

RQ1 *What is the effectiveness of LLMs to generate multiple-choice educational questions? What is the linguistic quality of the LLM-generated questions?*

To answer to this question, we design a controlled experiment in which we instruct the LLMs, through simple prompts, to generate multiple-choice questions from a given text context.

We then compare LLMs' performance with existing models, specifically fine-tuned for the task. Thus, we ask:

RQ2 *Is the LLMs' performance comparable with the state-of-the-art fine-tuned models for question generation?*

A positive answer to this question would encourage us and our industrial partner to invest time and effort in more sophisticated prompt engineering techniques with the aim of improving LLMs' effectiveness and generated questions quality without the need to conduct any fine-tuning tasks.

A. Empirical Setup

Dataset: Our evaluation relies on the SciQ [26] dataset, widely used for assessing EduQG models. SciQ contains 13,679 crowd-sourced science exam questions, divided into training, validation, and testing sets. We focus our evaluation on the testing set (1,000 questions) that makes possible a fair comparison with the baselines in RQ2. For each data point, SciQ includes: (a) *question*; (b) *correct_answer*; (c) list of *distractors*; and (d) *support_text*.

Prompts: We use the following two simple prompts in our empirical study, with the aim to instruct the LLMs to generate a multiple-choice question from the given support text.

– *Context_Only*: Given support text "support_text", create 1 expert level question with multiple choice answer from the text. Please also include the correct answer and 3 distractors.

– *Context_plus_answer*: Given support text "support_text", create 1 expert level question with multiple choice answer from the text, for which the correct answer is "correct_answer". Please, also create 3 distractors.

These prompts are adapted from the work of Fawzi et al. [6], where the authors showed that models tend to produce more concrete questions when the expected correct answer is provided as input, in addition to the support text. While Fawzi et al. [6] applied ChatGPT on just 9 cases from SciQ, this work provides an exhaustive experimental evaluation that provide concrete evidence on the potential of LLMs for the automatic question generation problem. It is worth remarking that we also include to the prompt, instructions regarding the format of the output, automatically generated by Pydantic Parser [2].

Selected LLMs: In total, our evaluation uses 16 LLMs, shown in Table I, including 7 models from Open AI, 4 from Meta Llama, 4 from Mistral AI, and 1 from DeepSeek.

Baselines: We use as a baseline the five fine-tuned models decribed in Section II, specialized in educational question

TABLE I: LLMs studied.

Open AI
gpt-4o, gpt-4o-mini, gpt-4-turbo, gpt-4, gpt-3.5-turbo-1106, gpt-3.5-turbo, gpt-4-0613
Meta Llama
Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct, Llama-3-70B-Instruct, Llama-3-8B-Instruct
Mistral AI
Mistral-7B-Instruct-v0.3, Mistral-Small-Instruct-2409, Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1
DeepSeek
DeepSeek-R1-Distill-Qwen-32B

TABLE II: Effectiveness of LLMs and the baselines.

Model	**Prediction Performance**					**Linguistic Quality**	
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	F1-score↑	Perplexity↓	Diversity↑
LLMs with Prompt: Context_plus_Answer							
Llama-3.1-70B-Instruct	50.43	42.51	38.04	34.99	59.61	12.42	80.79
GPT-3.5-Turbo-1106	49.67	41.87	37.39	34.34	57.66	9.52	82.54
Mixtral-8x7B-Instruct-v0.1	42.36	34.30	30.16	27.44	50.57	9.78	79.65
DeepSeek-R1-Distill-Qwen-32B	43.58	34.90	29.99	26.64	50.08	7.38	85.67
LLMs with Prompt: Context_Only							
Llama-3.1-70B-Instruct	31.61	23.47	19.91	17.95	36.77	7.28	79.62
GPT-3.5-Turbo-1106	30.91	22.95	19.49	17.59	35.43	10.69	79.51
Mixtral-8x7B-Instruct-v0.1	29.53	21.71	18.43	16.67	35.16	8.67	76.83
DeepSeek-R1-Distill-Qwen-32B	27.32	19.17	15.97	14.26	28.38	8.68	83.11
Baselines							
RLLM-EduQG	51.08	43.86	39.87	36.4	56.80	24.88	94.40
EduQG+	37.20	33.86	28.49	22.35	43.04	33.88	81.20
Leaf+	36.67	31.45	28.17	24.26	41.65	26.43	80.10
EduQG Large	29.19	21.69	18.03	16.76	33.18	34.36	74.90
EduQG Small	29.07	21.52	17.49	15.94	33.12	34.51	73.60
Leaf	27.07	20.22	17.17	16.46	30.90	30.82	73.50

generation, namely, Leaf [22], EduQG [5], and their variants Leaf+ and EduQG+, and RLLM-EduQG [11], that constitutes the current state-of-the-art. It is worth mentioning that we report the results presented in their corresponding papers.

Metrics: We rely on standard metrics widely used to evaluate two aspects of question generation models [5]: a) the prediction accuracy and b) the linguistic quality of the generated questions. To assess the predictive accuracy of the LLM-generated questions, w.r.t. the question in the ground-truth, we use the Bilingual Evaluation Understudy (BLEU) [14] score, which measures the syntactic similarity between the two sequences of tokens [14], and the METEOR [4] metric, that measures a word-to-word matching between the LLM-generated question and the human written reference. The computed matches and misses are then used to compute the precision, recall and the F1 score, that is the one reported.

To assess the linguistic quality of LLM-generated questions, we calculate their perplexity and diversity [25]. Perplexity is inversely related to the coherence of the generated text, and thus the lower the perplexity score, the higher the coherence of the generated question. The diversity score measures the number of distinct n-grams in the generated question (we use 3-grams as in [20]). Hence, larger diversity values, coupled with low perplexity, indicate the use of a richer vocabulary with grammatical precision. We reuse the same implementation of all the metrics provided by Bulathwela et al. [5].

Data Availability: In the following we discuss the results of the best performing models for each family, but all the generated questions and results are available at the following link: **https://github.com/rdegiovanni/LLMsForEduQG**.

B. Results RQ1: LLMs' Effectiveness

Table II summarizes the average of both performance and linguistic metrics obtained with the two prompt variations. Figures 1 shows the distribution of the F1 score and the BLEU-1 score obtained by the best performing models for each family. We can observe that the Llama-3.1-70B-Instruct model (and similarly, Llama-3-70B-Instruct) is the best performing in terms of the performance metrics. It reaches 59.61% of F1-score, almost 2% higher than the best GPT performing model (GPT-3.5-Turbo-1106 gets 57.66% F1-score, and GPT4o 55.93), and 10% higher than the best models from Mistral and DeepSeek families. GPT-4o-mini, had a similar performance as DeepSeek, obtaining a 50.39% of F1-score.

There are two key observations worth remarking out of these results. On the one hand, *open-weights and open-source* models obtained a very good and promising performance, even outperforming the GPT proprietary models. This evidence can help to increase and ease the adoption of open-source LLMs for Educational Question Generation in the near future. On the other hand, the prompt variant Context_plus_Answer that includes more precise information regarding the generative intention (e.g. the expected answer for the question to generate) is significantly more effective than general prompts (like the Context_Only variant). This observation should be taken into account in future prompt engineering in order to improve LLMs' effectiveness in educational question generation.

Regarding linguistic quality, we use as a baseline the perplexity score and diversity score, 18.74 and 82.40, respectively, of the ground truth questions from the SciQ dataset [5]. From Table II, we can see that LLM generated questions generally obtain a much better (lower) perplexity score and similar or even higher diversity score. This shows that LLM-generated questions use coherent language and are likely to be human-readable. In this regard, DeepSeek performed the best, with a perplexity score of 7.38 (2.54 times lower than ground truth) and a diversity score of 85.67 (3% higher than ground truth).

C. Results RQ2: Comparison with fine-tuned models

Table II also reports the prediction performance of all the specialized baseline models. We can observe that, even with basic prompts, all the LLMs obtain higher performance than Leaf, EduQG, and their variants. In addition, GPT models perform slightly better than the state-of-the-art RLLM-EduQG model, while Llama-3.1-70B-Instruct (similarly, Llama-3-70B-Instruct) is the best performing model, reaching almost 3% higher F1-score. This observation is in line with recent works [21], [23] that show that effective

TABLE III: LLMs' generated exact matches.

Model	**Num. of Exact Matches**		
	Question	Answer	Distractors
Prompt: Context_plus_Answer			
Llama-3.1-70B-Instruct	85	84	18
GPT-3.5-Turbo-1106	75	75	19
Mixtral-8x7B-Instruct-v0.1	49	47	9
DeepSeek-R1-Distill-Qwen-32B	28	28	10
Prompt: Context_Only			
Llama-3.1-70B-Instruct	23	21	5
GPT-3.5-Turbo-1106	22	20	4
Mixtral-8x7B-Instruct-v0.1	13	11	3
DeepSeek-R1-Distill-Qwen-32B	9	7	2

prompt engineering strategies on state-of-the-art LLMs can bring better results than fine-tuning the models from scratch, helping us to mitigate the effort in finding enough quality data and resources for performing the fine-tuning process.

Regarding the linguistic quality, LLMs obtained a much better (lower) perplexity score than all the baselines, while LLMs' diversity score is better than most of Leaf and EduQG variants, but worse (lower) than RLLM-EduQG model. While RLLM-EduQG can potentially produce richer questions than LLMs, the later can produce questions that are more cohesive and consistent than RLLM-EduQG. Overall, LLMs and the baselines have shown strong abilities to produce questions with very good linguistic quality.

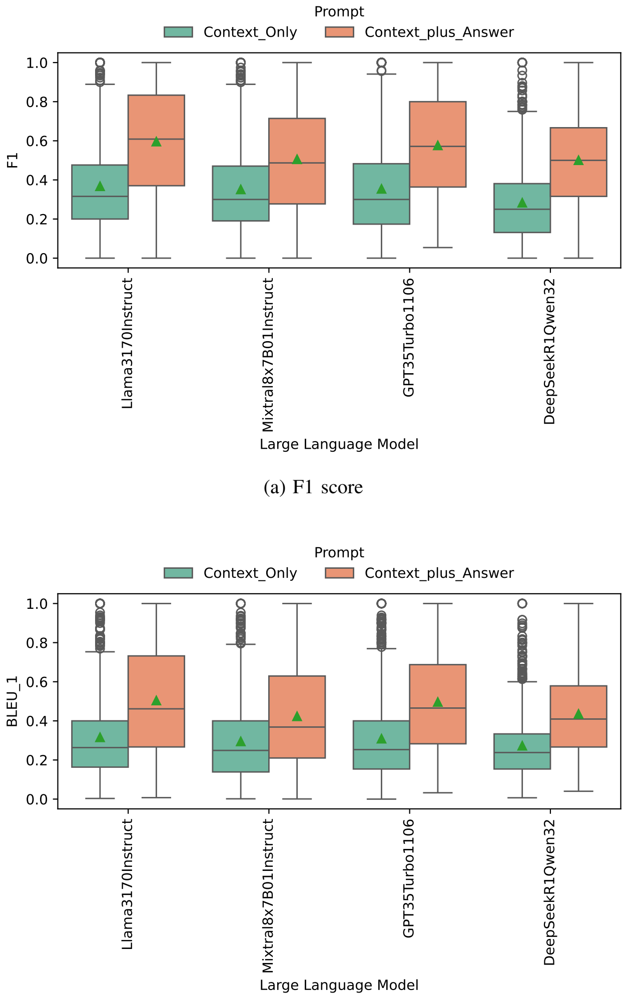
IV. Discussion and Lessons Learned

A. *Specific prompts improve the generation performance*

Since the experimental results are very promising, there are a few observations we would like to discuss to help future prompt engineering strategies. On the one hand, it is clear that the prompt Context_plus_Answer can generate questions that are more similar to the ground truth than prompt Context_Only. Table III shows the number of generated questions that exactly match with the ones in the ground truth (see Question column). We can observe that the prompt including the expected answer guides better the generative model and produces up to 4 times more questions that exactly matches with the ground truth. Hence, if the aim is to generate concrete and closed multiple-choice questions, the prompt should guide the LLMs with concrete keywords that may serve as the answer to the generated questions. On the other hand, questions generated just from the context (Context_Only) are not necessarily incorrect. Indeed, we inspected them manually and look alright, but these are open-questions, which can also be useful for a different scenario used in our experiments.

B. *Quality of the generated answer and distractors*

We now focus on the choices generated by the LLMs and analyze whether the generated answer and distractors match with the ones in the ground truth. Notice that we only consider the cases in which the generated question matches with the ground-truth, and report the results in Table III. We can observe that in most of the cases, whenever the generated question matches with the ground truth, the generated answer also matches with the ground truth. However, only the 25% of the generated distractors, on average, matches with the ones in the ground truth. Again, this does not mean that the remaining 75% of the generated distractors are not useful, but a more in-depth analysis, possibly requiring the assistance of the domain experts of our industrial partner, is required.



(a) F1 score

(b) BLEU-1 score

Fig. 1: LLMs' F1 score and BLEU-1 score on the SciQ dataset.

C. *Generalizability of the results*

Our experiments were based on the SciQ [26] dataset that includes questions mainly from Physics, Chemistry and Biology. To study the generalization of our results, we run the same experiments on 46 multiple-choice questions from the Canterbury Question Bank (CQB) [18], which target more theoretical Computer Science topics (our industrial partner plans to integrate programing exercises in the future). Table IV

TABLE IV: Evaluation on the Canterbury Question Bank [18].

Model	Prediction Performance		Linguistic Quality	
	BLEU-1↑	F1-score↑	Perplexity↓	Diversity↑
Prompt: Context_plus_Answer				
Llama-3.1-70B-Instruct	48.61	53.36	13.19	80.92
GPT-3.5-Turbo-1106	40.46	44.55	16.53	83.35
DeepSeek-R1-Distill-Qwen-32B	41.42	46.17	12.78	85.36
Mixtral-8x7B-Instruct-v0.1	40.58	44.24	9.77	81.23
Prompt: Context_Only				
Llama-3.1-70B-Instruct	37.97	38.10	12.81	81.99
Mixtral-8x7B-Instruct-v0.1	33.28	35.65	10.92	80.64
DeepSeek-R1-Distill-Qwen-32B	34.90	33.55	12.06	85.42
GPT-3.5-Turbo-1106	31.91	32.77	20.19	81.42

summarizes the prediction performance and linguistic quality of LLMs on the CQB dataset. We can observe that LLMs follow a similar trend as in the SciQ dataset, being Llama-3.1-70B-Instruct the best performing, and reaching higher predictions with the prompt Context_plus_Answer. Notice however that, when this prompt is used, the F1-score drops around 6%-8% w.r.t. the average F1-score on the SciQ dataset (e.g. Llama-3.1-70B-Instruct obtained 59.61% in SciQ vs 53.36% in CQB). While on the other hand, when the prompt Context_Only was used, LLMs' F1-score got improved (e.g. Llama-3.1-70B-Instruct obtained 36.77% in SciQ vs 38.10% in CQB), with the exception of GPT-3.5-Turbo-1106 model whose performance got reduced in 2%. The performance of Mistral and DeekSeek models follow similar trend as the one observed for Llama. Overall, despite minor changes in the prediction performance scores, LLMs show similar trends (ranking) than in the SciQ dataset, specially when the most effective prompt was used. Then, as it is reasonable to expect having better LLMs and prompts in the future, our observation suggest that the results of this work may generalize well to other datasets and domain.

D. LLMs' capabilities/limitations to produce structured output

It is worth remarking that most of the 16 LLMs included in our experiments correctly followed the instructions and produced structured outputs that we could parse and analyse, for instance, to compute all the metrics reported. We observed however that we failed in parsing ∼29%, ∼40% and ∼50% of the questions generated by GPT4o, Mistral-7B-Instruct-v0.2 and Mistral-7B-Instruct-v0.3, respectively, which can limit their adoption in practice. Moreover, we initially included other models in the experiments, namely, Google Flan, Google MT5, and Tiiuae Falcon, but were later discarded because most of the generated questions did not follow the correct structure to be parsed. While the generated text might be useful, this technical limitation was prohibited and had to exclude them from the evaluation and hence, it can also be a serious limitation of their adoption in practice.

V. Open Challenges and Roadmap

Advanced Prompting: The experimental results are promising, and more sophisticated prompt engineering techniques can potentially improve LLMs performance. We seek to explore Few-Shot and Chain-of-Thought Prompting [17], and Retrieval-Augmented Generation [12], that have been shown to be effective in improving related NLP-based tasks.

Contextual Information: As discussed with our industrial partner, the textual context might not be sufficient to determine the adequacy of the generated questions. Specific contextual information of the target audience, such as demographic information of the students and the expected difficulty, should be taken into account, and the questions should be adapted accordingly.

Supported by the company's domain experts, we are identifying the relevant dimensions we should take into account when designing our prompting strategies. We are in the process of defining a taxonomy that captures the target students metadata, that we can exploit to guide the LLMs to produce more student-specific questions.

Quality assessment of generated questions: Standardised exams have associated a diverse set of quality criteria that are relevant for our industrial partner. For instance, in multiple-choice questions, if the distractors are very similar to the correct answer, the question becomes ambiguous and unsuitable for assessment activities. We plan to use LLMs-as-a-Judge [8] strategies to implement a set of validators to identify issues in the generated questions, not complying with the expected quality properties.

Metamorphic testing to improve robustness: We plan to use metamorphic testing techniques [19] to study and improve the robustness of the designed LLM-based validators. Metamorphic transformations can be used to inject issues in the questions that should be identified by the validators, such as selecting a distractor as the correct answer. Other kinds of transformations, such as replacing a word by a synonym, should not change the semantic of the question, and can be used to assess LLMs robustness.

VI. Conclusion

This work provides concrete evidence on the basic LLMs' abilities to generate questions in educational contexts. It shows a general overview of the current state-of-the-art, integrating 16 LLMs from four different families, including commercial and open-source models. We observed that, even with very basic prompts, LLMs have a better performance than existing fine-tuned specialized models. We also showed that open-weights and open-source LLMs are as competitive as the commercial ones, which can be a point in favor to increase their adoption in practice, including our industrial partner. These results encourage us to explore more sophisticated prompt engineering strategies in the future, for which we have identified relevant open challenges that we are currently working on to improve the robustness and reliability of LLMs.

Acknowledgments

This work has been partially funded by the RDI Law project "Innovations for 21st Century Assessment Authoring" financed by the Luxembourg Ministry of the Economy, and the Luxembourg National Research Fund (FNR) PEARL program (grant agreement 16544475).

References

[1] Innovations for 21st century assessment authoring, financed by the luxembourg ministry of the economy. https://www.taotesting.com/blog/resources/innovations-for-21st-century-assessment-authoring-a-collaboration-between-oat-and-list/.

[2] Pydantic parser. https://python.langchain.com/v0.1/docs/modules/model_io/output_parsers/types/pydantic/.

[3] David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*, April 2023.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[5] Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. Scalable educational question generation with pre-trained language models. In *Artificial Intelligence in Education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings*, page 327–339, Berlin, Heidelberg, 2023. Springer-Verlag.

[6] Sahan Bulathwela Fares Fawzi, Sadie Amini. Small generative language models for educational question generation. In *NeurIPS'23 Workshop on Generative AI for Education(GAIED)*, 2023.

[7] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024.

[8] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[9] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Trans. Softw. Eng. Methodol.*, September 2024. Just Accepted.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025.

[11] Salima Lamsiyah, Abdelkader El Mahdaouy, Aria Nourbakhsh, and Christoph Schommer. Fine-tuning a large language model with reinforcement learning for educational question generation. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education*, pages 424–438, Cham, 2024. Springer Nature Switzerland.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[13] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020.

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

[17] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2024.

[18] Kate Sanders, Marzieh Ahmadzadeh, Tony Clear, Stephen H. Edwards, Mikey Goldweber, Chris Johnson, Raymond Lister, Robert McCartney, Elizabeth Patitsas, and Jaime Spacco. The canterbury questionbank: building a repository of multiple-choice cs1 and cs2 questions. In *Proceedings of the ITiCSE Working Group Reports Conference on Innovation and Technology in Computer Science Education-Working Group Reports*, ITiCSE -WGR '13, page 33–52, New York, NY, USA, 2013. Association for Computing Machinery.

[19] Sergio Segura, Gordon Fraser, Ana Belén Sánchez, and Antonio Ruiz Cortés. A survey on metamorphic testing. *IEEE Trans. Software Eng.*, 42(9):805–824, 2016.

[20] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation, 2017.

[21] Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks, 2023.

[22] Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. Leaf: Multiple-choice question generation. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 321–328, Berlin, Heidelberg, 2022. Springer-Verlag.

[23] Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R. Lyu. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, page 382–394, New York, NY, USA, 2022. Association for Computing Machinery.

[24] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. Software testing with large language models: Survey, landscape, and vision. *IEEE Trans. Softw. Eng.*, 50(4):911–936, February 2024.

[25] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. Towards human-like educational question generation with large language models. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 153–166, Cham, 2022. Springer International Publishing.

[26] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.