Towards Human-like Educational Question Generation with Small Language Models

Fares Fawzi, Sarang Balan, Mutlu Cukurova,
Emine Yilmaz, and Sahan Bulathwela

University College London, The United Kingdom.
{fares.fawzi.21,sarang.balan.20,m.cukurova,
emine.yilmaz,m.bulathwela}@ucl.ac.uk

Abstract. With the advent of Generative AI models, the automatic generation of educational questions plays a key role in developing online education. This work compares large-language model-based (LLM)systems and their small-language model (sLM) counterparts for educational question generation. Our experiments, quantitatively and qualitatively, demonstrate that sLMs can produce educational questions with comparable quality by further pre-training and fine-tuning.

Keywords: Question Generation · AI in Education · small Language Models.

1 Introduction

Large Language Models (LLM) have revolutionised educational applications with Artificial Intelligence (AI). Scalable educational question generation (EdQG) is a direct beneficiary of this trend. While recent studies use Model-as-a-Service (MaaS) products leveraging externally deployed LLMs (eg. ChatGPT) to carry out the educational question/quiz generation [7], such settings pose severe privacy, ethical and control-related issues. Model retraining can heavily affect model behaviour, compromising prompts and all downstream applications dependent on the MaaS LLM [14, 19, 15]. Limitations also arise during domain adaptation due to substantial training costs. Also, hosting LLMs on-premise is infeasible operationally and financially for the majority of education stakeholders. Small Language Models (sLMs), trained to excel in educational tasks, are a practical alternative that can unlock the quality of service without compromising control and stability. However, objectively comparing sLMs to LLM alternatives is a critical missing piece that this work attempts to address. We define sLMs as models that are easy to store, transfer and deploy ($\leq$ 250MB size) [9].

2 Related Work

In EdQG, state-of-the-art (SOTA) systems use pre-trained language models (PLMs) like Google T5 [17]. Recent EdQG research follows i) zero-shot prompt

engineering/tuning [8, 3] and ii) few-shot fine-tuning [4, 21]. Our work focuses on showing that fine-tuned sLMs can match the performance of LLMs on quantitative and human evaluations.

Recent work uses enormous LLMs that require significant computational power and expertise to train and maintain, including MaaS systems (e.g. ChatGPT [8]). MaaS API services carry the risk of undesirable changes in the behaviour of the host model, API usage limits and pricing changes - all of which pose different risks to the educators with little to no control over the models. Therefore, sLMs can be more desirable and safe in educational applications where the organisation owns and controls the language model (LM) with minimal expert and infrastructural costs. Recent works demonstrate how general-purpose sLM (T5-Small specifically) can be enhanced for EdQG through pre-training [4]. Our work extends their work by comparing the sLM's performance to LLM counterparts while assessing the human-readiness of sLM generations. This critical information affecting the adaptation of sLMs was not covered in [4]. We also measure the effects of post-grammar correction (GC) as sLMs fine-tuned for specific tasks (such as GC) can be used to improve LLM outputs [23].

Leaf (used in [4]), our baseline, is a SOTA LLM system that addresses EdQG by fine-tuning a pre-trained T5 PLM [17] with the SQuAD 1.1 dataset [18]. However, the SciQ dataset [25], a collection of 13,679 crowd-sourced scientific exam questions covering physics, chemistry and other sciences, is better suited for evaluating EdQG systems. [4] uses the S2ORC corpus with English scholarly abstracts [12] to make the model more suited for EdQG. We use EduQG proposed by [4] as the reference sLM in our experiments. Metrics such as BLEU, BERTScore, Human Ratings, Perplexity and Diversity are utilised [8, 24, 20, 13] to measure the quality of EdQG which are also used in this study.

Human evaluation is a reliable way to assess QG models and typical attributes such as fluency, relevance, answerability and usefulness are measured using Likert scales [8, 3]. In our study, we measure fluency, answerability and relevance. When collecting measurements, different prior works have used n-point Likert scales to rate the generations, with a 5-point Likert scale being the most common choice [11, 1]. We use a 5-point Likert scale from *strongly disagree* to *strongly agree*. We also use preference ratings to measure human preference for AI-generated questions (like [1, 3]).

3 Methodology

We aim to answer three main research questions.

- RQ1: Can sLM-based automatic grammar correction further improve EdQG?
- RQ2: How does sLM EdQG quality compare to general-purpose LLMs?
- RQ3: Are sLM generated questions humanly-acceptable?

3.1 Models, Datasets and Evaluation Metrics

We utilised two sLMs in experiments addressing RQ1: i) EduQG [4] which is based on the `T5-small` model (60.5M parameters) and ii) EduQG + a lightweight

RoBERTa-based GECToR model [16] (127M parameters) for grammar correction ($EduQG + GC$). In RQ2 experiments, we replicate Leaf [21], based on the `T5-base` (223M parameters), and use GPT3.5-based ChatGPT [1].

To reduce computational costs, a downsampled S2ORC dataset (2.1 million scientific abstracts) was used to pre-train a t5 model to create EduQG. The full SQuAD 1.1 dataset and the test set of the SciQ dataset were used for fine-tuning and evaluation respectively [4]. Furthermore, we randomly selected *9* SciQ contexts with the prompt "*Given text [context], create 5 expert-level questions with multiple choice answers from the text*" and selected the contexts where ChatGPT generates questions with the same answer as the SciQ dataset. Figure 1 illustrates this methodology.

Similar to [4], we use BLEU 1 through 4 (BL-1,...,BL-4) and F1-score (F1) to evaluate the predictive power of the models. We further use BERTScore [26], consisting of BERT Precision (B-Pr.), BERT Recall (B-Re.) and BERT F1 (B-F1.) to assess the semantic similarity of generations to the ground truth as the generative models may not use the same tokens and word order. Perplexity and Diversity are used to measure linguistic quality.

RQ3 was addressed through a questionnaire consisting of 2 parts. Part 1 is a pairwise preference task. Part 2 is a qualitative assessment task. The final part records demography and English fluency. In the Pairwise Preference Task (part 1 of the questionnaire), a pair of questions (A and B) were presented to the participant: i) a Teacher-generated ground truth(human-generated) and ii) an EduQG + GC model-generated (AI-generated) version of the same question. The ordering of the pair is randomised. The participants provide preferences for use in a teaching task based on a 5-level Likert scale (strongly prefer A, prefer A, no preference, prefer B, strongly prefer B). In the Qualitative Assessment (part 2 of the questionnaire), for each AI-generated question, 3 questions are asked about the level of i) Fluency, ii) Answerability and iii) Relevance. Again, a 5-level Likert scale is provided for all 3 questions with detailed descriptions of the definition of each aspect. The candidate questions used in this part were specifically selected to avoid overlap with items in Part 1 to prevent the learning effect and label leakage.

3.2 Experimental Setup:

The experimental setup to investigate RQ 1 and 2 is presented in Figure 1. The EduQG model, and its output through an sLM fine-tuned for English grammar correction ($EduQG + GC$), are analysed to answer RQ1. Leaf, ChatGPT and $EduQG + GC$ are compared to answer RQ2. Finally, the outputs from the $EduQG + GC$ sLM system are used for the user study (RQ3).

[1] https://chat.openai.com

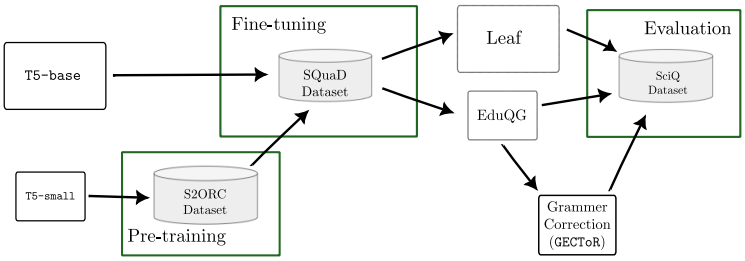

Fig. 1. Methodology for training and evaluating the models to answer RQ 1 and 2.

4 Results and Discussion

Table 1 shows how sLM-based EduQG, EduQG + GC (grammar corrected) and LLM-based systems Leaf and ChatGPT perform on the EdQG task. The perplexity calculation uses `TextDescriptives` [10] with the Spacy `en_core_web_lg` model as the reference PLM. Figure 2 further summarises the key results obtained from the user study.

Table 1. Top section: Comparison of predictive performance between leaf baseline (T5-base-based) and EduQG (T5-small-based sLM) on SciQ testset. **Bottom section**: Comparison of predictive performance between leaf baseline (T5-base-based), EduQG (T5-small-based sLM), and ChatGPT on 9 randomly selected contexts from SciQ testset. The best and second-best performance is indicated in **bold** and *italic* faces respectively.

	Predictive Performance								Language	
Model	BL-1	BL-2	BL-3	BL-4	F1	B-Pr.	B-Re.	B-F1	Perp.	Div.
Leaf LLM (†)	**0.9545**	**0.8176**	**0.6754**	**0.5737**	**0.6528**	**0.9279**	**0.9057**	**0.9165**	1.2942	0.7488
EduQG	0.9468	0.7750	0.6131	0.5016	*0.6044*	0.9145	0.8938	0.9039	**1.2675**	*0.7529*
EduQG + GC	*0.9470*	*0.7796*	*0.6202*	*0.5095*	0.6021	*0.9151*	*0.8944*	*0.9045*	*1.2813*	**0.7555**
Leaf LLM (†)	**0.7522**	**0.5450**	**0.3816**	**0.3080**	*0.4675*	*0.8995*	0.8636	0.8810	*1.3406*	*0.7503*
ChatGPT (†)	0.6071	0.4219	0.3146	0.2630	0.3941	0.8928	**0.8749**	*0.8836*	1.5001	**0.8200**
EduQG + GC	*0.7456*	*0.5018*	*0.3620*	*0.2997*	**0.4838**	**0.9064**	*0.8678*	**0.8865**	**1.2819**	0.7399

4.1 sLM vs. LLM-based EdQG systems (RQ 1 and 2)

Among the sLMs, Table 1 (top section) shows that the *EduQG + GC* model shows superior performance against EduQG, indicating the value addition of automatic post-grammar correction for this task (RQ1). The same section also highlights that EduQG systems based on sLMs perform similarly to the much larger ($\approx 4\times$) LLM-based Leaf counterpart. sLMs also outperform the Leaf baseline in perplexity and diversity. This is mainly because the sLMs are further pre-trained with scientific abstracts. This observation is very insightful as empirical evidence shows comparable performance in STEM-subject-related EdQG

can be obtained with significantly lightweight models. While sLMs are intriguing practically and scale-wise, results suggest that sLMs still struggle to capture grammatical structures fully, lending to their limited capacity. However, given that the grammar correction model itself is a sLM, the union of the 2 sLMs (EduQG and GECToR) is still significantly smaller than the larger baseline. Table 1 (bottom section) shows that the $EduQG + GC$ model is again comparable to the Leaf baseline while consistently outperforming ChatGPT outputs in the smaller dataset. While the ChatGPT experiment is smaller-scale ($n = 9$), this is promising evidence of the utility of sLMs in place of MaaS-based enormous LLM services like ChatGPT.

4.2 Human Evaluation (RQ3)

The participant set ($n = 9$) which is higher than the median of ($n = 3$) found in the literature [22]. The group consisted of 5 female (55.5%) and 4 male (44.5%) participants. While none of them were native English speakers, they all had post-secondary education to the Master's level at the minimum. 8 participants came from the 25-34 age bracket while the remaining one belonged to the 35-44 age bracket. All but two participants studied Mathematics, Biology, Chemistry, Physics and Earth Science in high school - which are the domains covered by the SciQ dataset.

The summary of results from the pairwise preference assessment study (part 1) is presented in Figure 2 (i). The majority of participants said that they either prefer AI-generated questions or have no preference between human or AI-generated questions for 12 out of 20 (60%) questions. This shows that the questions generated by the $EduQG + GC$ model are perceived to have equal or higher quality compared to human-created questions. The summary results from the qualitative assessment (part 2) are presented in Figure 2(ii). We observe the median scores for fluency, answerability and relevance factors are above 3, the centre of the 5-level Likert scale (1 to 5). This suggests that the participants express a positive sentiment regarding the quality of the AI-generated questions. Specifically, the fluency score is concentrated around very high values close to 5 suggesting the high linguistic quality of the generated questions. In comparison, answerability scores lie slightly lower with two outliers.

While the qualitative assessment highlights an above-average positive result, the generation quality has significant room for improvement. sLM models that generate education questions hold promise, yet the lack of overwhelming acceptance strongly suggests that the model outputs need to be improved significantly before any kind of deployment of sLMs for EdQG in a mainstream fashion.

4.3 Impact, Limitations and Future Research

Many recent works show zero-shot or prompt-tuned question generation to be operationally feasible using very large language models gated behind private APIs of MaaS services [7, 3]. Our results contribute to this topic as we demonstrate the utility of openly available sLMs to support EdQG. The proposed models are

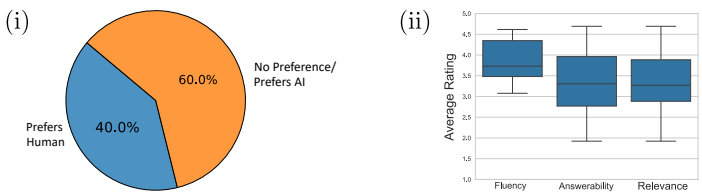

Fig. 2. Summary results from the user study.

very lightweight and open-source, giving the stakeholders full control and ownership, a critical feature for quality assurance of the downstream educational systems that rely on these models. Privately owned models carry less cybersecurity and data risk as all network and data interactions would occur within the organisation, as opposed to sending data to an external host. Additionally, the high power consumption needed to use LLMs marks a negative impact on its environmental sustainability. While the proposed models are not yet perfect, our results are positive and indicate that an educator can re-purpose these questions with minimum effort and time. When improved, educational questions can be generated at scale using the proposed model both for existing and new learning resources, adding more testing opportunities for learners/teachers.

Currently, an evident limitation of the EduQG model (even with grammar correction) from the results in section 4 is its inferior performance in comparison to LLMs. The proposed model still needs to be improved significantly to match the performance of LLMs. We hypothesise the sub-par performance attributes to 1) the model size, with 60M parameters and 2) training on 2.1 million out of 81 million available scientific abstracts. Our future work will aim to explore and unblock these bottlenecks. Furthermore, the statistical confidence of the ChatGPT experiment reported here is weak due to the small subset of data points that were used ($n = 9$) at this point. A larger scale comparison with more contexts (using the API) is necessary in the future to derive a better understanding of the sLM behaviour in comparison to ChatGPT. It is also possible to incorporate later versions of ChatGPT in future studies.

At last, we also need to be cautious to avoid the obvious pitfalls of such automatic systems. Intelligent QG models we build tend to exhibit the patterns in the data that we feed them. The pre-trained models we use as a foundation for building these sLMs are already trained with Internet data that is present with many biases. It is sensible to use post-processing tools to detect biases (e.g. [2]) and handle them before questions generated by these models are exposed to learners. Adaptation and assessing the usefulness of sLMs for cross-subject and cross-lingual question generation is another open research question that is under-explored at present.

5 Conclusion

In this work, we compare the performance of LLM-based QG models and sLM-based QG models in the context of EdQG. While the sLM models do not outperform their much larger counterparts, the results show that their generation capabilities are similar, and may be acceptable by humans, while the models being almost four times smaller. Reduced model sizes have significant advantages over larger language models in training and maintaining the models in-house, whilst retaining full ownership to be used in downstream educational services. This improves quality assurance as well as operational and capital costs by enabling complete control and oversight over their behaviours. We see our work being foundational to building a series of tools that can support educators with scalable personalised learning while scaling up question banks and knowledge bases in education [5]. The human-AI collaborative systems emerging initially can also produce valuable data that can be used to further fine-tune models. Ultimately, these models can be improved to the point where an intelligent tutor can create on-demand questions to verify a learner's knowledge state [6].

Acknowledgements This work is funded by the European Commission-funded projects "Humane AI" (grant 820437) and "X5GON" (grant No 761758). This work was also partially supported by the UCL Changemakers grant.

References

1. Amidei, J., Piwek, P., Willis, A.: The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In: Proc. of the 12th Int. Conf. on Natural Language Generation. ACL (2019)
2. Bai, Y., Zhao, J., Shi, J., Wei, T., Wu, X., He, L.: FairBench: A Four-Stage Automatic Framework for Detecting Stereotypes and Biases in Large Language Models. arXiv e-prints arXiv:2308.10397 (Aug 2023)
3. Blobstein, A., Izmaylov, D., Yifat, T., Levy, M., Segal, A.: Angel: A new generation tool for learning material based questions and answers. In: Proc. of the NeurIPS Workshop on Generative AI for Education (GAIED)
4. Bulathwela, S., Muse, H., Yilmaz, E.: Scalable educational question generation with pre-trained language models. In: Proc. of Int. Conf. on Artificial Intelligence in Education. pp. 327–339. Springer (2023)
5. Bulathwela, S., Pérez-Ortiz, M., Holloway, C., Cukurova, M., Shawe-Taylor, J.: Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. Sustainability **16**(2) (2024)
6. Bulathwela, Sahan and Pérez-Ortiz, María and Yilmaz, Emine and Shawe-Taylor, John: Power to the Learner: Towards Human-Intuitive and Integrative Recommendations with Open Educational Resources. Sustainability **14**(18) (2022)
7. Elkins, S., Kochmar, E., Cheung, J.C., Serban, I.: How teachers can use large language models and bloom's taxonomy to create educational quizzes. In: AAAI Conference on Artificial Intelligence (2024)
8. Elkins, S., Kochmar, E., Serban, I., Cheung, J.C.: How useful are educational questions generated by large language models? In: Proc. of Int. Conf. on Artificial Intelligence in Education. Springer (2023)

9. Fawzi, F., Amini, S., Bulathwela, S.: Small generative language models for educational question generation. In: Proc. of the NeurIPS Workshop on GAIED
10. Hansen, L., Olsen, L.R., Enevoldsen, K.: Textdescriptives: A python package for calculating a large variety of metrics from text. Journal of Open Source Software **8**(84), 5153 (2023)
11. van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., Krahmer, E.: Best practices for the human evaluation of automatically generated text. In: Proc. of the 12th Int. Conf. on Natural Language Generation. ACL (2019)
12. Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.: S2ORC: The semantic scholar open research corpus. In: Proc. of the Ann. Meet. of the ACL. Online (2020)
13. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Simplifying paragraph-level question generation via transformer language models. In: PRICAI 2021: Trends in Artificial Intelligence. Springer International Publishing (2021)
14. Loya, M., Sinha, D., Futrell, R.: Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In: Findings of the ACL: EMNLP 2023. pp. 3711–3716. ACL (2023)
15. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Proc. of the ACL (Vol 1: Long Papers). ACL (2022)
16. Omelianchuk, K., Atrasevych, V., Chernodub, A., Skurzhanskyi, O.: GECToR – grammatical error correction: Tag, not rewrite. In: Proc. of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 163–170. ACL, Seattle, WA, USA → Online (2020)
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
18. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proc. of the 2016 Conf. on EMNLP. ACL (2016)
19. Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A.: Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. arXiv preprint arXiv:2310.11324 (2023)
20. Ushio, A., Alva-Manchego, F., Camacho-Collados, J.: A practical toolkit for multilingual question and answer generation. In: Proc. of the 61st Annual Meeting of the ACL (Volume 3: System Demonstrations). pp. 86–94. ACL (2023)
21. Vachev, K., Hardalov, M., Karadzhov, G., Georgiev, G., Koychev, I., Nakov, P.: Leaf: Multiple-choice question generation. In: Proc. of the European Conf. on Information Retrieval (2022)
22. van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E.: Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech Language **67**, 101–151 (2021)
23. Vernikos, G., Brazinskas, A., Adamek, J., Mallinson, J., Severyn, A., Malmi, E.: Small language models improve giants by rewriting their outputs. In: Proc. of the 18th Conf. of the European Chapter of the ACL (Vol 1: Long Papers). ACL (2024)
24. Wang, Z., Valdez, J., Basu Mallick, D., Baraniuk, R.G.: Towards human-like educational question generation with large language models. In: Proc. of Int. Conf. on Artificial Intelligence in Education (2022)
25. Welbl, J., Liu, N.F., Gardner, M.: Crowdsourcing multiple choice science questions. In: Proc. of the 3rd Workshop on Noisy User-generated Text. ACL (2017)
26. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: Proc. of 8th Int. Conf. on Learning Representations. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr