Small Generative Language Models for Educational Question Generation

Fares Fawzi[1], **Sadie Amini**[2] **and Sahan Bulathwela**[1,3]
[1] Department of Computer Science [2] The Bartlett School of Architecture
[3] Centre for Artificial Intelligence
University College London
The United Kingdom
{m.bulathwela}@ucl.ac.uk

Abstract

The automatic generation of educational questions will play a key role in scaling online education, enabling self-assessment at scale when a global population is manoeuvring their personalised learning journeys. This work compares the predictive performance of foundational large-language model-based systems and their small-language model counterparts for educational question generation. Our experiments demonstrate that small language models can produce educational questions with comparable quality by further pre-training and fine-tuning while producing very lightweight models that can be easily trained, stored and deployed.

1 Introduction

Digital learning resources such as Massively Open Online Courses (MOOC) and Open Educational Resources (OER) are abundant, but they often lack associated questions that enable self-testing and skill verification [5, 7] once the learning resources are consumed. Generating scalable educational questions is crucial for democratising education [6]. While existing large language models (LLM) are used for question generation, their utility in education has only been explored recently. While LLMs bring opportunities, they demand expensive training and maintenance costs (expert man-hours and infrastructure) that the majority of stakeholders in education find challenging to meet. If Small Language Models (sLM) were as good as LLMs for this task, they could mitigate the challenges.

This work investigates the feasibility of using sLMs in place of LLMs in providing comparative AI-generated outputs in learning activities. In this work's context, sLMs are defined as significantly smaller models that have much fewer parameters making them lightweight to store, transfer and deploy (Under 250MB in file size). Specifically, we study the case of *educational question generation* where 1) context is provided alone as input and 2) the answer is provided with the context.

2 Related Work

Automatic question generation (QG) involves creating valid and coherent questions based on given sentences (and optionally, desired responses) [26]. By leveraging question-answering datasets, neural models can generate questions using both the context and expected response, ensuring high-quality questions. However, this approach often relies on an additional system to identify relevant responses [20], limiting its real-world applicability. Alternatively, QG models can be trained to depend solely on the context, allowing the creation of questions that belong to a specific type [24] for the document, paragraph, or sentence [11, 8]. This work focuses on both these task settings.

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

2.1 Pre-trained Language Models (PLMs) for Educational QG

In the field of educational neural QG, state-of-the-art (SOTA) systems leverage pre-trained language models (PLMs) such as GPT-3 [3] and Google T5 [17]. These models, pre-trained on massive text corpora, enable zero-shot question generation without additional training. Recent research has investigated and demonstrated the potential for generating educational questions using 3rd party hosted foundational models such as ChatGPT [22, 2, 9].

Leaf, a cutting-edge question generation system, fine-tunes an LLM for the question and multiple-choice distracter generation [21]. It uses the SQuAD 1.1 dataset [18] to train its question generation component by fine-tuning a pre-trained T5 model [17]. To evaluate the quality of generated questions, various metrics such as BLEU, ROUGE, METEOR, F1-Score, Human Ratings, Perplexity, and Diversity are utilised [2, 22, 13]. While the results presented in these works are promising, the models they used are very big and require enormous computational power and expertise to train and maintain. Due to this reason, many recent works are limited to leveraging LLMs that are hosted by third parties (e.g. ChatGPT [9, 10]. However, the API approach risks an uncertain behaviour for the educational tools built on top of the model as the end-user (host organisation) has no control over the behaviour of the LLM, especially in the case where the host model is periodically retrained and the new checkpoint can behave significantly differently to its predecessor. Hence, the behaviour of the downstream application is severely compromised. Therefore, sLMs can be more desirable and safe in such applications as the application owner can own the LM with minimal expert and infrastructural costs.

Recent work has shown how general-purpose sLM (T5-Small specifically) can be further enhanced for educational question generation through pre-training with scientific text giving birth to EduQG models [15, 4]. This work attempts to compare their generation capacity to LLMs (T5-base) in the question generation task. This pushes us in the direction of quantifying the opportunity cost of opting for sLMs over LLMs.

2.2 Related Datasets

For question generation (QG) and question-answering (QA) datasets, [26] offers a comprehensive review. The Leaf system, our baseline, is designed for educational purposes by fine-tuning the T5 model using the SQuAD 1.1 dataset [18]. However, this dataset is less suited for evaluating educational QG capabilities. In contrast, SciQ [23] is a collection of 13,679 crowd-sourced scientific exam questions covering physics, chemistry, and other sciences that is more suitable for evaluating educational question generation. We use SQuAD 1.1 and SciQ datasets for fine-tuning and evaluation respectively. Prior work uses the S2ORC corpus comprising 81.1 million English scholarly abstracts [12] to make the model more suited for educational QG [4]. We use the same approach here to make the models more education-focused.

3 Methodology

The primary objective is to compare the relative performance of the education-focused sLM proposed in [15, 4] to SOTA LLMs used for educational QG. We identify three key research questions in this direction.

- **RQ1:** How does the education-specific sLM perform in comparison to a larger general-purpose LM in educational QG when the answer **is/ is not provided** as input?
- **RQ2:** How does an education-specific sLM's output questions compare to a SOTA prompt-based system like ChatGPT?
- **RQ3:** For the contexts tested with chatGPT, Can the sLM-generated questions be accepted by human evaluators?

3.1 Models

We use a similar methodology used in [4]. As the primary LLM baseline model, we replicate the leaf system [21], where the `T5-base` (Model size: 223M parameters, $\approx$ 900 MB file)[1] model [17] is

[1] `https://huggingface.co/t5-base/tree/main`

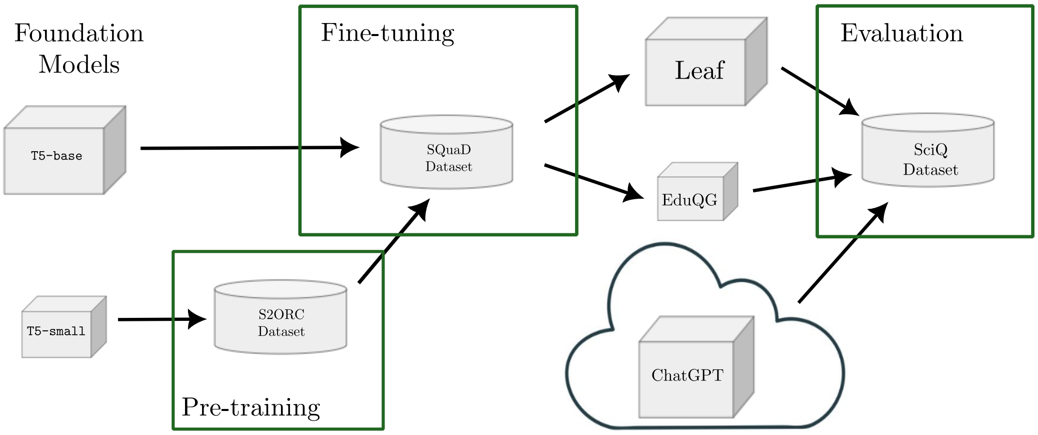

Figure 1: Methodology for training and evaluating the baseline Leaf model (above, for RQ1), EduQG (middle) and ChatGPT (cloud, for RQ2) models.

fine-tuned for question generation. Furthermore, we also use ChatGPT API as many works use this model for educational QG [22, 9]. When using ChatGPT, the prompt used is "Given text <context>, create 5 expert level questions with multiple choice answers from the text". As for the sLM, we replicate the EduQG model [4] with the `T5-small` (Model size: 60.5M parameters, ≈ 240 MB file)[2] model as its foundational LM.

3.2 Datasets and Evaluation Metrics

The datasets mentioned in section 2.2 are used. To create the EduQG model, 2.1 million scientific abstracts in the S2ORC dataset are used for pre-training. Both the baseline leaf model and EduQG models are fine-tuned with the SQuAD 1.1 dataset for QG. As per [4], the test set of the SciQ dataset is used to assess QG capabilities. Similar to [4], BLEU 1-4 and F1-score are used for the evaluation of the models. The BLEU 1-4 is used to assess the precision of the generation while the F1-score evaluates the balance between the precision and the recall. [19, 16] Both are commonly used metrics and used in prior work [2, 13, 18]. Our experimental setup is presented in figure 1.

4 Results and Discussion

Tables 1 and 2 show how the trained LMs (sLM-based EduQG, LLM-based Leaf and ChatGPT) performed on educational QG tasks in both task settings where 1) the answer was not provided as part of the input, and 2) when the answer was provided as part of the input.

4.1 Performance of sLM vs. LLM-based QG systems (RQ1)

Table 1: Comparison of predictive performance between leaf baseline (T5-base-based) and EduQG (T5-small-based). The better performance is indicated in **bold**.

Task	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	F1-Score
Without Answers	Leaf LLM Baseline	**0.9285**	**0.7738**	**0.6251**	**0.5171**	**0.5843**
	EduQG sLM	0.9231	0.7531	0.5994	0.4885	0.5741
With Answers	Leaf LLM Baseline	**0.9545**	**0.8176**	**0.6754**	**0.5737**	**0.6528**
	EduQG sLM	0.9499	0.8051	0.6609	0.5616	0.6516

The results in table 1 show that the EduQG models that are based on sLMs perform very close to the much larger ($\approx 4x$), more complex to train and maintain LLM-based Leaf counterparts in both task settings. In a computer that has 2 x Nvidia Titan RTX GPUs, fine-tuning the sLM-based model takes 8 minutes on average *per epoch* compared to the LLM-based model which takes 40 minutes

[2] `https://huggingface.co/t5-small/tree/main`

on average. This indicates a significant reduction in computation. This means that sLMs can be trained to perform well in question generation systems in both settings where the answer can be provided with the input or otherwise (RQ1). This observation is very insightful for AI in education practitioners as it gives empirical evidence that comparable performance in STEM-subject-related question generation can be obtained with significantly lightweight models by further pre-training them on the right type of data.

Another key observation from table 1 is that the predictive performance of models trained and prompted *with* the answer is much higher compared to the models that do not use the desired answer as part of the input prompt. However, this observation can be due to two reasons. More specifically,

1. The model may have a disadvantage in forming a good question when an expected answer is not provided with the prompt
2. The model may be generating valid and coherent questions, yet disadvantaged in generating the *expected* question in the labelled dataset due to the absence of the answer

The random examples presented in prior work [4] suggest that the questions generated by T5-small-based QG models are coherent in general. However, this aspect should be studied in detail with extensive evaluations.

4.2 Comparing to ChatGPT (RQ2) and Human Readiness (RQ3)

A small subset of (9 contexts) from the SciQ dataset was also processed using ChatGPT [3] to generate questions (as per bottom part of figure 1). The results in table 2 show the predictive performance of the models used for RQ1 experiment and ChatGPT. The obtained results indicate that the locally trained Leaf baseline outperforms the ChatGPT-generated questions in terms of BLEU and F1 metrics across the board. It is also observable that the sLM-based EduQG model shows slightly superior performance to ChatGPT in unigram-based performance metrics (specifically, BLEU-1 and F1 score). This observation is an indication of higher word overlap between the sLM-generated question to the human expert-generated question although the word ordering is not consistently aligned with the human label as much as ChatGPTs. This can suggest 1) more grammatical errors in sLM model or 2) it is using a different style of language.

In order to assess the human-readiness of question generation (RQ3), the questions generated from the EduQG model were examined. The following are the questions generated by sLM-based EduQG model for the 9 contexts that were tested with ChatGPT.

1. What are Compounds capable of donating electron'd compounds called?
2. What is the first diploid cell of a new organism?
3. What is the backbone of a vertebra?
4. What is the most easily oxidized magnesium?
5. What is the rate of reaction when reactants are higher?
6. What is the proximal portion of endoderm tissue that extends anteriorly from foregut?
7. What are mammals called that lay eggs instead of giving birth to an infant or embryo?
8. What is the process of heat transfer?
9. What is equal to the cross product?

The generated questions are grammatically accurate for the most part without the need for any human intervention. This result demonstrates that even the sLM-based QG model can work well in a human-in-the-loop system that assists teachers and teaching assistants in generating questions for educational materials while the models do most of the generation. The generated output is compelling in trialling the proposed models in a human-facing system. But, these results are very limited in statistical power to make confident conclusions. A follow-up user study would enable us to clearly verify the utility of the sLM-based models.

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

Table 2: Comparison of predictive performance between leaf baseline (T5-base-based), ChatGPT and EduQG (T5-small-based) on a subset of contexts from the SciQ dataset. The best and second best performance is indicated in **bold** and *italic* faces respectively.

Task	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	F1-Score
Baseline	Leaf LLM Baseline	**0.7275**	**0.5054**	**0.3536**	**0.2909**	**0.4749**
LLMs	ChatGPT API	0.6071	*0.4219*	*0.3146*	*0.2631*	0.3941
sLM	EduQG	*0.6357*	0.3877	0.2544	0.2076	*0.4059*

4.3 Impact and Limitations

Many works in the past have shown how zero-shot question generation is operationally feasible using very large language models gated behind an API from a large corporation (Model-as-a-Service architecture) [22, 9]. However, our result contributes to this topic as we demonstrate the utility of openly available sLMs (in our case, T5-small-based EduQG) to support educational QG. We intentionally use the *T5-Small* model that has 60.5M parameters in comparison models such as GPT-3 XL that has 1.3B parameters [3] to show that relatively small models can be trained with domestic hardware to create SOTA educational QG capabilities. The proposed models are very small and lightweight, giving the stakeholders the potential to have full control and ownership, a critical feature for quality assurance of the downstream educational systems that rely on these models (contrary to having no control when relying on an API-managed LLM through a third party that can change their model over time). While the proposed models may not be perfect, the quality of AI-generated questions in prior work [4] indicates that a teacher or an educator can re-purpose these questions with minimum effort and time. There is promise to build Human-in-the-loop systems that can support educators to scale their learning systems to more learners. Educational questions can be generated at scale using the proposed model both for existing and newly created learning resources, adding more testing opportunities for learners/teachers to use when needed. We see our work being foundational to building a series of tools that can support educators with scalable/personalised assessments while leveraging tools to scale up question banks and knowledge bases in educational topics. Ultimately, we have the opportunity to improve these models to the point where an intelligent tutor can rely on them to create on-demand questions to verify a learner's knowledge state with no human intervention (e.g. in knowledge tracing).

We 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. We need to be mindful that we take rigorous steps to validate the datasets to be ethically and pedagogically sound. Putting emphasis on quality assurance of the training data will help us to build ethical, unbiased QG models that can benefit all learners equally. However, the pre-trained models we use as a foundation for building these sLMs are already trained with Internet data that is present with many of these biases. It is sensible to use post-processing tools to detect biases (e.g. [1, 25]) and handle them to make sure the questions generated by these models are responsibly exposed to learners. Another gap in this work is the lack of human evaluation of the AI-generated questions. While offline evaluation on labelled datasets is useful, having teachers and learners evaluate and compare human vs. AI-generated questions will provide much more insightful findings into the pedagogical quality of the questions that can improve this line of research in the future. Assessing the usefulness of the proposed sLMs to provide cross-domain and cross-lingual questions is also an open research question that is under-explored at present.

5 Conclusion

In this work, we strive to compare the generation performance of LLM-based general-purpose QG models and sLM-based QG models for educational use cases. Specifically, their question generation capabilities in the context of STEM-subject QG. While the sLM models do not show the ability to outperform their much larger counterparts, the results show that the generation capabilities are very similar while the latter family of models are almost 4 times smaller than their larger counterparts. This reduction of model size has advantages in training and maintaining these models in-house rather than relying on third-party services. Having maintainability of models in-house also gives the host organisation full ownership of the models that they use in downstream educational services. This degree of control also safeguards the host organisation by giving the ability for quality assurance as the organisation has full control over retraining the models and testing their new behaviours before rolling out. While the quantitative metrics reflect slightly inferior generation capabilities, eyeballing a sub-sample of generated questions from the sLM-based QG model shows that the generated questions are still high in quality and are useable for educational question generation in the wild.

Based on the several limitations of the current work pointed out in section 4.3, running human evaluations on the generated questions is one of the key next steps to understanding the human readiness of these systems further. While algorithmic methods have emerged recently to assess more complex aspects such as bias [1] and answerability [14] of generated questions, they can be understood better with user studies. Furthermore, identifying new datasets to improve cross-domain (beyond STEM) and cross-lingual capabilities of the proposed models is also beneficial to widen their impact. Our subsequent work will focus on this aspect.

Acknowledgements This work is also partially supported by the European Commission-funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437), EU Erasmus+ project 621586-EPP-1-2020-1-NO-EPPKA2-KA and the EPSRC Fellowship "Task Based Information Retrieval" (grant EP/P024289/1). This research is conducted as part of the X5GON project (`www.x5gon.org`) funded by the EU's Horizon 2020 grant No 761758. This work was also partially supported by the UCL Changemakers AI Co-creator grant.

References

[1] Y. Bai, J. Zhao, J. Shi, T. Wei, X. Wu, and L. He. FairBench: A Four-Stage Automatic Framework for Detecting Stereotypes and Biases in Large Language Models. *arXiv e-prints*, page arXiv:2308.10397, Aug. 2023.

[2] S. Bhat, H. Nguyen, S. Moore, J. Stamper, M. Sakr, and E. Nyberg. Towards automated generation and evaluation of questions in educational domains. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 701–704, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] S. Bulathwela, H. Muse, and E. Yilmaz. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer, 2023.

[5] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources. In *AAAI Conference on Artificial Intelligence*, 2020.

[6] S. Bulathwela, M. Pérez-Ortiz, C. Holloway, and J. Shawe-Taylor. Could ai democratise education? socio-technical imaginaries of an edtech revolution. In *Proc. of the NeurIPS Workshop on Machine Learning for the Developing World (ML4D)*. arXiv, 2021.

[7] 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.

[8] X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] S. Elkins, E. Kochmar, I. Serban, and J. C. Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

[10] R. Goran and D. Abed Bariche. Leveraging gpt-3 as a question generator in swedish for high school teachers, 2023.

[11] H. Guo, R. Pasunuru, and M. Bansal. Soft layer-specific multi-task summarization with entailment and question generation. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The semantic scholar open research corpus. In *Proc. of the Ann. Meet. of the ACL*, Online, 2020.

[13] L. E. Lopez, D. K. Cruz, J. C. B. Cruz, and C. Cheng. Simplifying paragraph-level question generation via transformer language models. In D. N. Pham, T. Theeramunkong, G. Governatori, and F. Liu, editors, *PRICAI 2021: Trends in Artificial Intelligence*, pages 323–334, Cham, 2021. Springer International Publishing.

[14] A. Mohammadshahi, T. Scialom, M. Yazdani, P. Yanki, A. Fan, J. Henderson, and M. Saeidi. RQUGE: Reference-free metric for evaluating question generation by answering the question. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada, July 2023. Association for Computational Linguistics.

[15] H. Muse, S. Bulathwela, and E. Yilmaz. Pre-training with scientific text improves educational question generation (student abstract). In *AAAI Conference on Artificial Intelligence*, 2023.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, 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] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 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, Nov. 2016. Association for Computational Linguistics.

[19] A. B. Sai, A. K. Mohankumar, and M. M. Khapra. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2), jan 2022.

[20] L. J. Tamang, R. Banjade, J. Chapagain, and V. Rus. Automatic question generation for scaffolding self-explanations for code comprehension. In *International Conference on Artificial Intelligence in Education*, pages 743–748. Springer, 2022.

[21] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, and P. Nakov. Leaf: Multiple-choice question generation. In *Proc. of the European Conf. on Information Retrieval*, 2022.

[22] Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk. Towards human-like educational question generation with large language models. In *Proc. of Int. Conf. on Artificial Intelligence in Education*, 2022.

[23] J. Welbl, N. F. Liu, and M. Gardner. Crowdsourcing multiple choice science questions. In *Proc. of the 3rd Workshop on Noisy User-generated Text*. ACL, Sept. 2017.

[24] X. Wu, N. Jiang, and Y. Wu. A question type driven and copy loss enhanced frameworkfor answer-agnostic neural question generation. In A. Birch, A. Finch, H. Hayashi, K. Heafield, M. Junczys-Dowmunt, I. Konstas, X. Li, G. Neubig, and Y. Oda, editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 69–78, Online, July 2020. Association for Computational Linguistics.

[25] W. Zekun, S. Bulathwela, and A. S. Koshiyama. Towards auditing large language models: Improving text-based stereotype detection. In *Proc. of the NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*, 2023.

[26] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng. A review on question generation from natural language text. *Trans. on Information Systems*, 40(1):1–43, 2021.