Small Generative Language Models for Educational Question Generation

Fares Fawzi[1], **Sadie Amini**[2], **Sahan Bulathwelaa**[1,3]

[1] Department of Computer Science, [2] The Bartlett School of Architecture,
[3] Centre for Artificial Intelligence

University College London

{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.

Introduction

Digital learning resources, such as Massively Open Online Courses (MOOCs) and Open Educational Resources (OERs) often lack associated questions that enable self-testing and skill verification [2, 5] after the learning resources are consumed. Generating scalable educational questions is a crucial step towards democratising education [4]. While the use of Large Language Models (LLM) has been explored for generating educational questions, their expensive training and maintenance costs pose challenges. This work explores the feasibility of using Small Language Models (sLM) as a smaller alternative to LLMs in *educational question generation* where 1) context is provided alone as input and 2) the answer is provided with the context.

Related Work

- Automatic question generation task settings
  - Question generation using both context and expected response [12]
  - Question generation using only the context [16, 8, 6]
- Educational neural Question Generation
  - Zero-shot pre-trained language models (PLMs), *Google T5* [10]
  - Third party hosted foundational models, *ChatGPT* [14, 1, 7]
  - Fine-tuning LLM on question and multiple-choice distracter generation
    - Leaf, fine-tuned a pre-trained T5 model on question generation [13]
  - Pre-training sLM to be enhanced for educational question generation
    - EduQG, T5-small pre-trained on scientific text [9, 3].

Related Datasets

- SQuAD 1.1 dataset [11], *Less suited for educational question generation.*
- SciQ [15], *More suitable for evaluating educational question generation*

Methodology

The primary objective is to compare the relative performance of the education-focused sLM proposed in [9, 3] to SOTA LLMs used for educational QG. We identify three key **research questions**:

- **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?

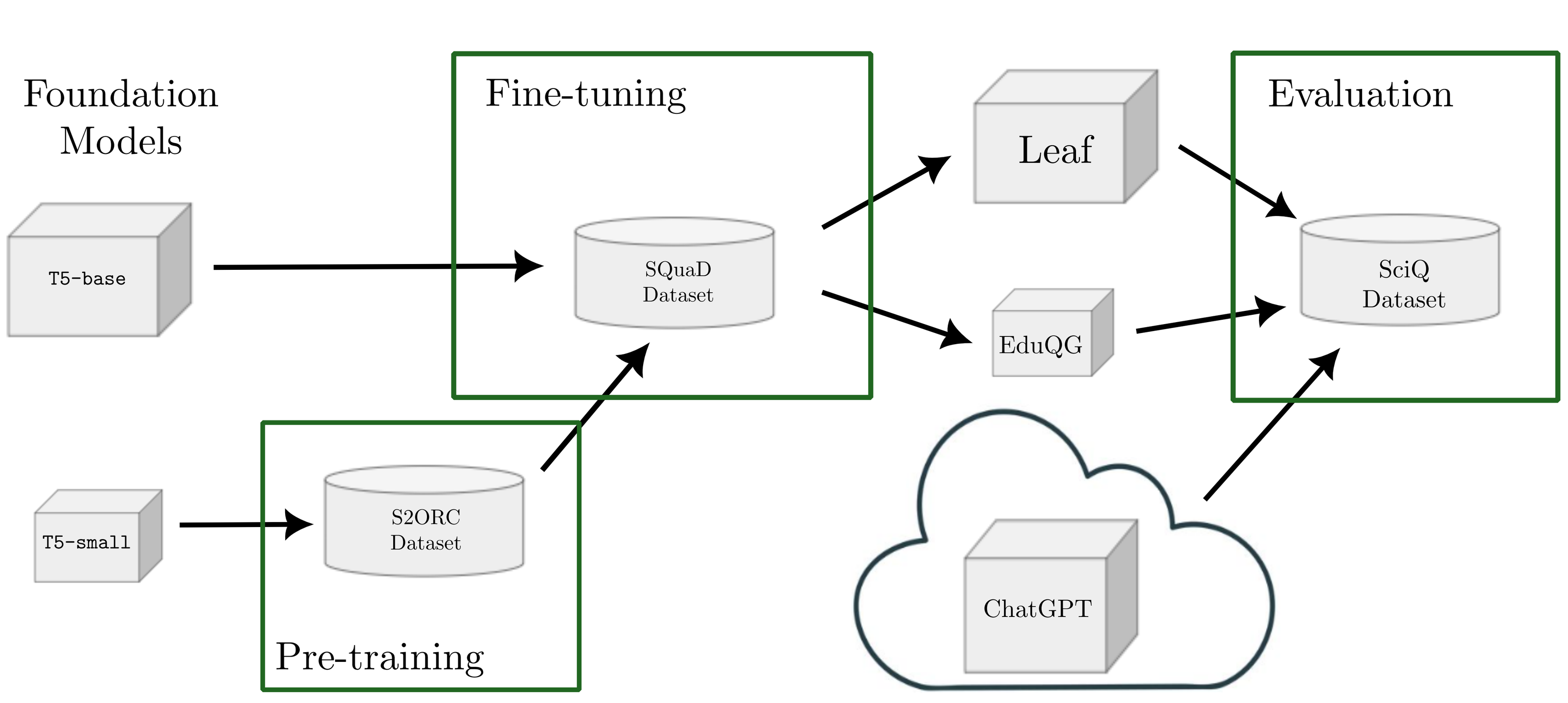

Figure 1: Methodology for training and evaluating the baseline Leaf model (above, for RQ1), EduQG (middle) and ChatGPT (cloud, for RQ2) models.

Results

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

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*

Conclusion

- Compared the generation performance of LLM-based general-purpose QG models and sLM-based QG models for educational use cases.
- The generation capabilities of sLM are very similar to models that are 4 times larger.
- Running human evaluations on the generated questions is one of the key next steps.
- Identifying new datasets to improve cross-domain (beyond STEM) and cross-lingual capabilities of the proposed models is another aspect that subsequent work will focus on.

Acknowledgements

This work is also 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) and EU Erasmus+ project 621586-EPP-1-2020-1-NO-EPPKA2-KA.

References

[1] Shravya Bhat, Huy Nguyen, Steven Moore, John Stamper, Majd Sakr, and Eric Nyberg. Towards automated generation and evaluation of questions in educational domains. In Antonija Mitrovic and Nigel 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.

[2] 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.

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

[4] Sahan Bulathwela, María Pérez-Ortiz, Catherine Holloway, and John 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.

[5] 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.

[6] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In Regina Barzilay and Min-Yen 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.

[7] Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

[8] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Soft layer-specific multi-task summarization with entailment and question generation. In Iryna Gurevych and Yusuke 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.

[9] Hamze Muse, Sahan Bulathwela, and Emine Yilmaz. Pre-training with scientific text improves educational question generation (student abstract). In *AAAI Conference on Artificial Intelligence*, 2023.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter 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.

[11] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy 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, November 2016. Association for Computational Linguistics.

[12] Lasang J Tamang, Rabin Banjade, Jeevan Chapagain, and Vasile Rus. Automatic question generation for scaffolding self-explanations for code comprehension. In *International Conference on Artificial Intelligence in Education*, pages 743–748. Springer, 2022.

[13] Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. Leaf: Multiple-choice question generation. In *Proc. of the European Conf. on Information Retrieval*, 2022.

[14] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. Towards human-like educational question generation with large language models. In *Proc. of Int. Conf. on Artificial Intelligence in Education*, 2022.

[15] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proc. of the 3rd Workshop on Noisy User-generated Text*. ACL, September 2017.

[16] Xiuyu Wu, Nan Jiang, and Yunfang Wu. A question type driven and copy loss enhanced frameworkfor answer-agnostic neural question generation. In Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Kenneth Heafield, Marcin Junczys-Dowmunt, Ioannis Konstas, Xian Li, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 69–78, Online, July 2020. Association for Computational Linguistics.