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 consuming very lightweight models that can be easily trained, stored and deployed.

Introduction
Digital learning resources, such as Massive 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 dictator generation
  - Leaf, fine-tuned a pre-trained TS 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.
- SQuAD [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 question generation. 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 (closed, for RQ2) models.](image-url)

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Answer</td>
<td>Leaf LLM Baseline</td>
<td>0.7275</td>
<td>0.5054</td>
<td>0.3536</td>
<td>0.2399</td>
<td>0.8093</td>
</tr>
<tr>
<td>With Answer</td>
<td>Leaf LLM Baseline</td>
<td>0.7551</td>
<td>0.6179</td>
<td>0.5171</td>
<td>0.3578</td>
<td>0.9491</td>
</tr>
</tbody>
</table>

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

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

**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 "Hume 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**


**Small Generative Language Models for Educational Question Generation**

Fares Fawzi1, Sadie Amini2, Saham Bulathwela1,3
1 Department of Computer Science, 2 The Bartlett School of Architecture, 3 Centre for Artificial Intelligence

University College London

{f.m.bulathwela}@ucl.ac.uk