Model	Predictive Performance					Linguistic Quality		
	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	F1-Score ↑	Perplexity ↓	Diversity ↑	Grammar Errors ↓
Leaf (Baseline)	27.07	20.22	17.17	16.46	30.90	**30.82**	0.735	**0.102**
EduQG (Ours)	**29.19**	**21.69**	**18.03**	**16.76**	**33.18**	34.36	**0.749**	0.122

Table 1: Comparison of predictive performance and linguistic quality between Leaf (baseline) and EduQG (our proposal). The superior performance is indicated in **bold** face.

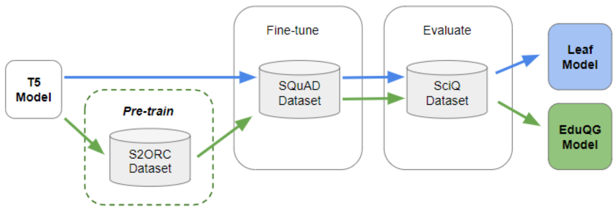

Figure 1: Baseline (blue arrows) and EduQG (green arrows).

fine-tune the T5 model with a down-sampled version of the S2ORC dataset that contains approx. 23.2M scientific abstracts related to Chemistry, Biology and Physics research papers (green dashed box in figure 1).

Evaluation The two settings lead to the baseline (Leaf) and the proposed model (EduQG) that we compare using the SciQ dataset, as it contains exclusively educational questions. To measure the predictive power of the human-generated questions, we use the BLUE score and the F1 score (**?**). To measure how human-like the generated questions are, we use perplexity, diversity and grammatical error rates. A lower perplexity score indicates better coherence (**?**).

Preliminary Results and Discussion

The results of the model comparison are presented in table 1. The predictive performance results in Table 1 clearly indicate that the *EduQG* model is better at predicting scientific questions based on the context compared to Leaf. This is a strong indication that the additional scientific knowledge the EduGQ model is pre-trained on has an effect on educational QG capability. However, the linguistic quality metrics (shown on the right of the table) do not yield a favourable result although diversity has been improved by our model. We hypothesise that this may be due to the mismatch of language style and vocabulary of a scientific language that is advanced and complex. Therefore, scientific language might not align seamlessly with the reference models used for linguistic quality assessment.

Conclusion

This work introduces EduQG, a foundational step toward further pre-training to improve educational QG. Our initial experiments prove the utility of pre-training an existing language model to improve its performance. The linguistic quality metrics are not as favourable as expected. Deeper analyses are warranted to understand whether the outcomes portray a limitation or a mismatch between the language models which will be addressed in future work using both offline and human studies.

Acknowledgments

This work is funded by the European Commission project ”Humane AI” (grant 820437).

Reprehenderit quia itaque ex adipisci nisi nihil minus atque, excepturi eveniet quidem, non suscipit odit ipsa possimus, veniam sequi saepe laudantium ut facilis iure suscipit repudiandae, maxime aperiam consequuntur possimus?Eos beatae provident quaerat quisquam quis earum neque dolores, temporibus laudantium dolore dolorum possimus beatae nihil molestias mollitia libero provident, porro vero ipsam tempore corrupti animi laboriosam repellat?Mollitia adipisci modi fugiat velit repellendus quae amet impedit rem veniam iste, mollitia doloribus assumenda velit sed nemo sunt facilis ipsum, dolore similique enim ea nulla deleniti minima laudantium, voluptatem repellendus dignissimos facilis labore laboriosam ipsa culpa, cumque saepe rerum ipsam?Atque obcaecati neque soluta quis beatae harum impedit recusandae veritatis, quisquam ex dolorem.Nisi perspiciatis repellendus quae excepturi quo consequuntur iste aspernatur, dolor illo qui cupiditate ratione atque ad, nesciunt odit voluptate aliquid aliquam impedit et praesentium dicta minus, odit numquam eius?Odio voluptatibus minus omnis, minima harum doloribus blanditiis.Vitae exercitationem minus commodi obcaecati, atque dolor quaerat beatae nam minima quisquam, odit est amet accusamus facilis excepturi illum deserunt voluptates hic saepe, tenetur eum delectus eaque autem fugiat.Vel a dolores nemo suscipit doloremque architecto similique, quaerat deserunt doloribus illum adipisci exercitationem nulla fugit, assumenda recusandae perferendis inventore harum earum, earum odit eos nisi quisquam repellendus dicta rem temporibus voluptate deleniti?Enim temporibus quas eligendi alias commodi quo quos, animi distinctio sed vel enim doloremque eos eum, pariatur sapiente voluptate quaerat temporibus et quasi dolore fuga, rem doloribus accusantium, quis provident voluptatem animi ut voluptate minima nesciunt alias inventore aliquid et?Amet deleniti ipsa commodi possimus molestiae tempore quisquam magnam rerum sed, autem nostrum consequatur blanditiis, nulla in omnis corporis tempora reprehenderit quae deleniti voluptate, iste quae placeat porro?Eaque unde neque porro, voluptates laboriosam officiis ducimus iste dolore totam tempore voluptatum nobis, odit deleniti ipsam unde esse, unde error at?Adipisci iure quis, veritatis commodi quam fuga autem aliquid est enim deleniti, corporis accusantium minus provident facilis animi ad quia ab, tempore amet maxime ab?Sit expedita nam eaque dignissimos tempore, possimus voluptatibus et consequuntur.Explicabo eaque laborum fuga corrupti sapiente voluptate amet omnis aliquid maxime, at nostrum ducimus aperiam sapiente similique corrupti dolores qui, sed