

DiffiQG: Difficulty-Controlled Question Generation Using Small Language Models

Sahan Bulathwela^{1,2}, Wilson Lee², Kalindu Kumarasinghe³ and Sarang Balan²

¹Centre for Artificial Intelligence, University College London, The United Kingdom

²Department of Computer Science, University College London, The United Kingdom

³Independent Researcher

Abstract

Educational Question Generation (EQG) can reduce teacher workload and support adaptive learning, yet most existing systems lack reliable control over question difficulty grounded in pedagogical frameworks or depend on computationally expensive large language models that pose affordability and ownership complexities, limiting accessibility. This paper presents DIFFIQG, a difficulty-controlled question generation framework that explicitly conditions generation on Bloom’s Taxonomy using lightweight, fine-tuned transformer models. We benchmark multiple Bloom’s classification strategies and identify a high-performing classifier (80.5% accuracy) to serve as an automatic evaluation oracle. Publicly available Bloom-annotated datasets are curated and augmented with generated instructional contexts to enable supervised conditional fine-tuning of Flan-T5 models. Experimental results show that a Flan-T5-Base model achieves strong difficulty control (91.3 % Bloom alignment accuracy) while maintaining competitive linguistic quality. Higher-order cognitive levels are generated reliably, while mid-range levels remain challenging due to overlapping conceptual boundaries. Human evaluation confirms strong agreement with automatic difficulty metrics. Overall, this work demonstrates that pedagogically grounded difficulty-controlled EQG can be achieved without large-scale models, paving the way to scalable and privacy-preserving personalised tutoring and assessment generation.

Keywords

Question Generation, Small Language Models, Question Difficulty, Bloom’s Taxonomy, Adaptive Learning

1. Introduction

Modern educators face escalating workload demands, with assessment design constituting a significant portion of instructional preparation time. Educational Question Generation (EQG) has emerged as a promising NLP application with the potential to enhance adaptive learning systems and reduce teacher burden with the use of generative AI (GenAI) [1, 2].

The effectiveness of educational assessments depends critically on question alignment with pedagogical objectives. Intelligent Tutoring Systems (ITSs) require mechanisms to adapt exercise difficulty to students’ knowledge levels, yet current EQG approaches typically lack inherent mechanisms for fine-grained difficulty control, introducing inefficiencies and limiting scalability. While transformer-based models (e.g. T5 [3]) have improved question fluency, they generate questions without explicit difficulty control. Existing approaches either employ proxy metrics disconnected from pedagogical frameworks or rely on computationally expensive Large Language Models (LLMs), limiting deployment in resource-constrained settings.

Contribution: This work develops a method for scalable difficulty-controlled Educational Question Generation (EQG) while enabling accessibility in low-resource educational environments. We (1) identify/develop a highly accurate difficulty classifier (80.5%) using the Bloom’s Taxonomy to proxy question difficulty, (2) fine-tune a small language model (sLM) for difficulty-controlled EQG (91.3% difficulty alignment), and (3) run a detailed analysis on this sLM’s behaviour on specific cognitive difficulty levels.

SLM4ED’26: The 1st Workshop of Small Language Models for Education (SLM4ED), June 28, 2026, Seoul, Republic of Korea

✉ m.bulathwela@ucl.ac.uk (S. Bulathwela)

🆔 0000-0002-5878-2143 (S. Bulathwela)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1.1. Problem Statement

A learner ℓ has consumed learning resources containing the knowledge context c . The teacher’s (or an ITS’) goal is then to generate a question $\hat{q}_b$, where $\hat{q}_b$ is an educational question at the Bloom’s taxonomy category b ($b \in \{\text{remember, understand, apply, analyse, evaluate and create}\}$). $\hat{q}_b$ is a sequence of tokens, $q_b \in \{w_1, \dots, w_{|q_b|}\}$ of length $|q_b|$. This task requires that 1) the question is contextually relevant to the context c , while closely aligning with Bloom’s level b . The probability $p(q_b|c, b)$ incorporates the coherence and relevance of each token in the sequence, rendering the generation process highly sensitive to both the context and the difficulty level. This task can be mathematically defined to identify the optimal question $\hat{q}_b$ that maximises the conditional probability as per equation 1.

$$\hat{q}_b = \arg \max_{q_b} p(q_b|c, b) = \arg \max_{q_b} \sum_{i=1}^{|q_b|} \log p(w_i|c, b, w_1 \dots w_{i-1}) \quad (1)$$

where, $p(q_b|c, b)$ denotes the conditional probability that also depends on the tokens $w \in q_b$.

2. Related Work

Question Generation and Difficulty Control: Recent advances leverage pre-trained language models in EQG, with models like T5 [3] demonstrating substantial improvements from classical machine learning methods. For constrained generations, topic-controlled EQG [4] and difficulty-controllable methods [5] have been attempted, while difficulty-controlled EQG remains underexplored, specifically with sLMs. Our focus is on filling this gap.

Difficulty in EQG is a loaded definition, operationalised through different lenses such as textual specificity, multi-hop reasoning demands, answer complexity and other computational methods [6, 7]. While insightful, these proxies of difficulties do not stem from established pedagogical frameworks, posing challenges to link them to formal teaching strategies. While frameworks like Bloom’s taxonomy successfully map cognitive complexity of learning artefacts (questions, learning objectives etc.), very little is done to incorporate linking Bloom’s taxonomy to EQG, although such trends have emerged recently from LLM space [8]. Like those works, we too use Bloom’s taxonomy to represent difficulty in this work, but focus on training sLMs rather than exploiting LLMs.

Bloom’s Taxonomy in Educational Assessment: Bloom’s Taxonomy leads literature as a widely applied framework for pedagogically classifying learning objectives and, by extension, assessment difficulty. The revised taxonomy [9] structures cognitive processes hierarchically: Remember, Understand, Apply, Analyze, Evaluate, and Create. While classical/neural machine learning-based Bloom’s taxonomy classifiers have emerged (e.g. [10], [11]), recent efforts leverage Bloom’s Taxonomy with LLMs to generate questions at specified difficulty levels [12, 13, 14]. However, much of this progress depends on computationally expensive proprietary models.

Scalability and Ethical Considerations in Educational AI: Many EQG systems require substantial computational resources or cloud-based infrastructure, creating accessibility barriers and privacy concerns in low-resource settings [15, 16], surfacing critical deployment barriers. sLMs offer an alternative paradigm to this challenge, promoting equitable access, lower barriers to local deployment and ownership of data and systems, leading to sustainability. Inspired by recent results showing the ability to train sLMs for topic-controlled EQG [4, 17, 18], we explore extending the scope of sLMs’ capability to control difficulty.

3. Methodology

Our work is driven by *three* prime research questions.

RQ1: Can a finetuned sLM-based Bloom’s taxonomy classifier outperform prompted LLMs?

RQ2: Can sLMs be trained for Bloom taxonomy-controlled EQG?

RQ3: Does the final EQG model perform well across all Bloom’s categories?

As our RQs entail creating both a classification and a generation model to attack the sub-tasks within the overarching aim, a phased experimentation approach is used, comprising 3 phases.

Phase 1: Developing Bloom’s level classifier for difficulty estimation (RQ1); **Phase 2:** Developing difficulty-controlled question generator (RQ 2 & 3); and **Phase 3:** Evaluating the difficulty-controlled generator (RQ 3). Transitioning through the phases, the best-performing model developed in each phase is used as an input to the next phase.

3.1. Models

We develop a distinct set of classification and generation models needed for the different phases of the work. Pedagogical alignment, computational efficiency and low-resource deployment feasibility were prioritised during model selection.

Bloom’s Taxonomy Classifiers (Phase 1) We experimented with an encoder-only sLM (very cheap to train and host) vs. foundational LLMs (requiring no training and is cheap to access through a hosted API). The models considered are:

Encoder-only Classifier (66M parameters), *DistilBERT* [19] is a lightweight competitive sLM base-model for multi-class classification. The model was fine-tuned for multi-class classification over six Bloom’s categories using cross-entropy loss.

To benchmark against contemporary instruction-tuned foundational LLMs, we evaluated *Gemini 2.5 Flash* (commercial, estimated 5B parameters), *Llama3-8B* (open, 8B parameters), and *Llama3-70B* (open, 70B parameters) configured as zero-shot and few-shot classifiers[20]. These models were prompted with explicit Bloom’s Taxonomy definitions and, where applicable, exemplar question–label pairs. All LLM classifiers produced categorical outputs constrained to the six Bloom’s levels. Inspired by Elkins et al. [21], we designed a prompt to provide a single context paragraph with one and five examples (single- vs multiple-shot) for in-context learning. The structure of the prompt used is as follows:

Classify the following question based on Bloom's Taxonomy using the provided context.

Bloom's Taxonomy Categories:

1. remember: Recall facts and basic concepts.
2. understand: Explain ideas or concepts.
3. apply: Use information in new situations.
4. analyse: Draw connections among ideas.
5. evaluate: Justify a stand or decision.
6. create: Produce new or original work.

Examples:

Context: Climate change refers to long-term shifts in temperatures and ...

Question: What are the human activities that contribute to climate change?

Classification: Remember

...

DiffiQG Models (Phase 2) For question generation, we selected instruction-tuned transformer models from the Flan-T5 family[22] which we finetune. To , assess model efficiency, we experiment with two model sizes, **1) Flan-T5-Base** (250M parameters, referred to as base hereafter) and **2) Flan-T5-Small** (80M parameters, referred to as small hereafter.) T5 family of models have performed well in EQG (both constrained [4] and unconstrained [3]) making it a suitable model family for this work.

Following the conventional prompt design template for training FLAN-T5 model, we constructed a prompt that can exploit the pre-trained embeddings of tokens for instruction understanding. The prompt template used is:

TASK: Generate a question based on the {BLOOM_LEVEL} and {CONTEXT}

3.2. Datasets and Data Curation

Our datasets for both classification and generation tasks require that the examples have 1) explicit Bloom’s Taxonomy annotations across all six cognitive levels, and 2) sufficient linguistic diversity.

3.2.1. Dataset for Classifier Training (Phase 1)

We combine the Gani et al. [11], Yahya et al. [23] and Sangodiah et al. [24] to combine all unique examples to create a **Combined Classification Dataset (CCD)** for training the sLM classifier. The Combined Classification Dataset, contains 2,521 examples, where 80% of the data is used for training, and the remainder for validation.

3.2.2. Dataset for training the Difficulty-controlled Question Generator (Phase 2)

We identify 2 potential datasets that can be combined for this task, namely, [12] and 2) Nicyscaria [13] datasets.

BloomLLM Dataset [12] is a 687 expert authored question dataset from postgraduate AI and machine learning with Bloom’s level annotations. All labels are expert annotated and reviewed with a balanced class distribution, making it a high quality reference set for training and evaluation.

nicyscaria Dataset [13] on the otherhand, is a 2,550 example LLM-generated question set for graduate level data science domain. Each question is annotated by 2 experts and only examples with inter-rater agreement is retained for quality assurance. This leads to 1,834 examples we use in our work. There is class imbalance in this dataset.

Handling Context Synthesis and Class Imbalance: Neither of the above datasets contains the source context (c in sec. 1.1, the paragraph the question is based on) creating a critical limitation. As our instruction tuning method requires the context, we use Gemini Flash 2.5, a high quality, commercial foundational model to generate five shot prompt-combining each question-topic pair, adapting the template designed for Bloom level classification in phase 1.

Due to the class imbalance primarily present the nicyscaria dataset, we oversample minority class observations in the combined dataset (by randomly duplicating examples in the dataset) to match the largest class (*understand*, $n = 458$).

Combined Generation Dataset (CGD): The retained datasets were merged into a unified corpus of 2,521 question–Bloom level pairs. Stratified sampling was applied to produce an 80:10:10 train/validation/test split, preserving class distributions across all subsets.

This synthesised augmentation transforms the dataset into (c, b, q_b) triplets, enabling supervised conditional generation consistent with Eq. 1. From now on, we refer to this dataset as **Combined Generation Dataset**.

3.3. Evaluation Metrics

The evaluation framework should allow assessing the *classification quality* of the classifier in Phase 1 while both the *linguistic quality* and the *pedagogical alignment* with the intended Bloom’s cognitive

level of the generation models in phase 2. As no single metric sufficiently captures educational text quality, we adopt a complementary set of automatic and human evaluation measures.

Bloom’s Level Classification Metrics (Phase 1): Classification accuracy is the primary metric used here. This metric tells us the agreement between the predicted Bloom’s category and the actual label allowing us to pick the best model.

Linguistic Quality Metrics (Phase 2): *BLEU* scores [25] measuring n -gram overlap between generated and reference questions (BLEU-1 capturing lexical adequacy to BLEU-4 reflecting longer phrase consistency) is used. While limited for open-ended generation, BLEU remains informative for surface-level fidelity in controlled EQG settings [4]. *METEOR* [26] accounts for synonymy and recall, providing a more flexible similarity measure suited to paraphrastic variation in question generation.

Perplexity serves as a proxy for fluency and coherence, with lower values indicating more predictable and well-formed output [3].

Difficulty Control and Pedagogical Alignment Metrics (Phase 2 & 3): *Bloom’s Classification Accuracy* is used to measure the difficulty level of questions based on Bloom’s taxonomy category. While human expert judgements are the gold standard, our hypothesis is that if the model generates a question conditioned on Bloom’s category, the generated question would return the respective Bloom’s level when classified using a reliable Bloom’s category classifier. In this case, we use the best classifier identified in phase 1 to verify if the generated question $\hat{q}_b$ belongs to the prescribed Bloom’s category b .

3.3.1. Human Evaluation

While expensive, human annotations still enrich the reliability of our results. We employ two postgraduate students (studying AI) as human evaluators for classifying a small sample of generated questions into Bloom’s difficulty levels. They independently annotate Bloom’s levels using standard taxonomy definitions. Inter-rater reliability was measured using Cohen’s Kappa (κ). *Human–AI Agreement*, representing the alignment between human annotations and automatic Bloom classifications, was analysed to further reinforce the reliability of the AI-based pedagogical alignment metric.

3.4. Overall Experimental Design

The experimental design is structured into three sequential phases, each targeting a specific subset of the research questions and collectively forming a unified evaluation pipeline (RQ3).

In phase 1, we use the sLM identified for fine-tuning the Bloom’s category classifier along with several LLMs with multiple prompting techniques to identify the most reliable difficulty classifier. The Combined Classification Dataset is used for finetuning the sLM. We use a subset of the Combined Generation Dataset (CGD) for an unbiased comparison of classification performance between the finetuned and the prompt-driven LLMs. Using the CGD for evaluation allows us to 1) make a fair comparison between the fine-tuned sLM classifier and the prompt-driven LLM classifiers, as neither uses the evaluation data source during training, and 2) determine how reliable classification metrics would be if we were to later train generation models on the CGD dataset in phase 2. The best performing model is promoted to Phases 2 and 3.

In phase 2, we take the CGD dataset described in sec. 3.2.2 to finetune the FLAN-T5 small and base models. We then evaluate the models using the CGD test set. **1) Base** and **2) Small** models are compared in this phase. Evaluation metrics, BLEU-1, BLEU-4, Perplexity and Bloom’s classification Accuracy described in 3.3 are used for evaluation.

In phase 3, the Cognitive-Level Performance Analysis (RQ3) is conducted based on the best generation model from phase 2. The final phase analyses model performance at individual Bloom’s difficulty categories to identify systematic strengths and weaknesses across cognitive levels. The generation model is scrutinised for class-level performance using post-generation, per-class classification accuracy,

which also allows studying emerging confusion patterns, particularly among adjacent cognitive levels (e.g., Understand, Apply, Analyse).

4. Results and Discussion

In this section, we present the results of the experiments run in the 3 phases to discuss findings.

4.1. Best Performing Bloom Category Classifier (RQ1)

Classification accuracy obtained by both finetuned and LLM-based models used in phase 1 shows that the Gemini-based classifier with single-shot prompting, presenting one example for each Bloom’s category from the same context paragraph, achieved the highest accuracy at 80.5%, exceeding human evaluation (76%) and outperforming DistilBERT (73%) and Llama3-70B (73%).

Contrary to expectations, the LLM-prompted models (not post-trained for this specific task) significantly outperform the finetuned DistilBERT model. This is interesting as fine-tuned sLMs (even when they are significantly smaller) tend to outperform general-purpose LLMs that are not finetuned for the specialised tasks (e.g. [17] from EQG itself). However, the expected phenomenon is observed for the Llama3-8B model (Accuracy at ~54%), which is modest in size compared to the 70B model.

Multi-context prompting (where 5 examples for each Bloom’s difficulty level were provided) followed a similar trend (Gemini and Llama 70B outperforming sLM counterpart) but performed worse in comparison to the single-shot approach. This suggests that excessive provision of examples can confuse the model rather than clarifying decision boundaries.

4.2. Overall EQG Model Performance (RQ2)

The results obtained from developing Flan-T5 model variants are reported, breaking them into linguistic quality (in tables 1 and 2) and difficulty alignment (in tables 1 and 3).

Based on table 1, the Base version (250M parameters) reports the better linguistic quality and Bloom taxonomy classification accuracy. The results indicate that within sLMs, the larger model with more parameters has significant performance gains. While generated questions achieve moderate overlap with reference questions (METEOR ~0.39, well below the theoretical maximum of 1.0), the low perplexity score indicates fluent, coherent language. The randomly selected examples in table 2 further strengthen the evidence that the generated questions demonstrate rich linguistic structure while also referring to action words linked to Bloom’s levels.

The human evaluation (in table 3) validates automatic metrics we have derived from the Gemini Flash LLM-based classifier, with human accuracy aligning closely with Bloom’s Classification Accuracy ($\kappa > 0.84$ across both setups, indicating reliable inter-rater agreement). The superiority of the base model is evident here as well. This is also a side effect of the better Bloom’s type accuracy of the base model.

4.3. Performance Across Bloom’s Cognitive Levels (RQ3)

Table 4 provides a class-level breakdown of classification accuracy. In summary, the Base model naturally leads in terms of performance except for a handful of classes while scoring ≥ 0.93 on 4/6 classes. **Evaluate** and **Create** levels achieved near-perfect accuracy (0.93-1.00) and this is primarily due to these classes having distinct linguistic patterns (e.g., ‘evaluate,’ ‘design,’ ‘develop’).

Table 1

Overall Performance Metrics for Fine-tuned Flan-T5 Models for Base and Small Model Versions

Model	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	Perplexity ↓	Bloom’s Accuracy ↑
Base	0.371	0.116	0.392	45.36	0.913
Small	0.343	0.097	0.368	48.15	0.715

Table 2

Randomly selected example outputs for the 4 different model setups FLAN-T5 (1) Base and (2) Small for two randomly selected Bloom’s taxonomy categories, Create and Evaluate classes.

Level	Model	Generated Question
Create	Base	Design a sequence-to-sequence transformer model to summarise a text document. Outline the metrics you would use to assess the model’s accuracy.
	Small	Create a chain-of-thought prompt for a student population of Indian students who seek educational resources in various subjects and languages. Include examples of prompting techniques that could be used to guide the model’s ability to break down multiple-lingual queries and present it in a structured, step-by-step manner.
Evaluate	Base	Evaluate the performance of a machine learning algorithm used for spam detection on a dataset on Kaggle.
	Small	Evaluate the effectiveness of a word embedding classifier in the classification of news articles in Indian languages. How would you evaluate the performance of a word embedding classifier in real-world contexts?

Table 3

Human Evaluation of Fine-tuned Flan-T5 Models

Model	Accuracy			Cohen’s κ
	AI	Human 1	Human 2	
Base	0.913	0.94	0.92	0.87
Small	0.715	0.73	0.75	0.84

Table 4

Bloom’s Classification Accuracy breakdown by Bloom’s Taxonomy Level

Model	Remember	Understand	Apply	Analyze	Evaluate	Create
Base	0.89	0.96	0.70	0.98	1.00	0.93
Small	0.65	0.80	0.70	0.20	0.97	1.00

By contrast, the **Apply** level remains challenging for the Base model (0.70 accuracy), with generated questions frequently confused with Understand or Analyse levels. Looking at the Small model results in table 4, it is evident that it has further challenges with the Analyse level, where the accuracy drops to 0.2.

Examples of misclassified Apply questions reveal the subtlety: questions asking students to “*explain how a mechanism works*” and “*given a confusion matrix, explain how to use it*” were generated as Apply-level. But they are classified as Understanding. The distinction between explaining a concept (Understand) and applying it to solve a problem (Apply) depends on implicit task complexity, not explicit lexical markers. While pre-trained transformer-based models are better at manoeuvring such nuances in comparison to classical machine learning representations, the subtleties between subsets of these levels may still make T5 models struggle, especially when the parameter size is limited.

4.4. Beneficial Applications and Limitations:

Our work demonstrates that significantly lightweight (Under 250M parameters) sLMs can be finetuned to perform complex tasks like difficulty-controlled EQG. The approach shows promise in scenarios where 1) finetuning with niche/proprietary data on specialised domains is essential, 2) contents of learning resources are required to stay within local infrastructure (due to copyright/privacy reasons), and 3) organisations wish not to face changing behaviour of cloud-based LLMs across evolving model checkpoints that might spawn service obligations to downstream services. Given the quality of the generations and the accuracy of targeted generation, these models form a foundational step toward creating intelligent tutors [27] that can offer ad hoc, personalised testing opportunities to individual learners with richer pedagogical alignment, where the tutor can subscribe to specialised micro-skills. The models may also be used to build human-AI systems that generate question banks at scale.

The current model is strictly tested on English generation on primarily AI/ML domain-based datasets. In that respect, the generalisation of quality and alignment capabilities to other domain areas and languages is untested. A critical comparison of the proposed methods with a prompted foundational LLM-based difficulty-controlled question generator is also absent in this work. While the sLMs generate reasonable questions, whether they outperform frontier LLMs is yet to be tested.

Another noteworthy realisation is that LLMs were still required to create the question-generating sLMs. As per the experiments, the Gemini 2.5 Flash model (est. 5B parameters) was utilised in the context generation and evaluation steps due to its superior performance. While the final generation models are sLMs, it is fair to acknowledge LLMs played a vital role in creating the final sLMs.

5. Conclusion

This research demonstrates that Bloom’s Taxonomy can be effectively integrated into lightweight transformer-based EQG systems, achieving 91.3% accuracy in difficulty control through systematic dataset curation, data synthesis, and fine-tuning. However, challenges remain in generating pedagogically accurate questions for mid-range cognitive levels (Apply, Analyse). These difficulties likely reflect genuine ambiguities in how these constructs are operationalised within authentic educational contexts. For educational contexts seeking scalable, privacy-preserving, and pedagogically aligned question generation, this work provides a practical foundation. The Flan-T5-Base model offers sufficient performance for deployment while maintaining accessibility characteristics essential for equitable educational technology.

A few limitations are identified, suggesting that future work should focus on multi-domain and multi-language generalisation, more nuanced modelling and using more sophisticated training methods like contrastive training.

Acknowledgments

This work is co-funded by the European Commission’s projects “Teacher-AI Complementarity (TaiCo)” (Project ID: 101177268), “Humane AI” (Grant No. 820437) and “X5GON” (Grant No. 761758) and the UCL Computer Science Strategic Research Fund.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for drafting content, grammar checking, and spelling checking. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] M. Giannakos, R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, S. Järvelä, M. Mavrikis, B. Rienties, The promise and challenges of generative ai in education, *Behaviour and Information Technology* (2024). doi:10.1080/0144929X.2024.2394886.
- [2] M. Heilman, N. A. Smith, Good question! statistical ranking for question generation, in: R. Kaplan, J. Burstein, M. Harper, G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 609–617.
- [3] S. Bulathwela, H. Muse, E. Yilmaz, Scalable educational question generation with pre-trained language models, in: *International Conference on Artificial Intelligence in Education*, Springer, 2023, pp. 327–339.

- [4] Z. Li, M. Cukurova, S. Bulathwela, A novel approach to scalable and automatic topic-controlled question generation in education, in: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025, pp. 148–158.
- [5] T. Goto, Y. Tomikawa, M. Uto, Enhancing diversity in difficulty-controllable question generation for reading comprehension via extended t5 (2024).
- [6] A. Faizan, S. Lohmann, Automatic generation of multiple choice questions from slide content using linked data, in: *Proceedings of the 8th international conference on web intelligence, mining and semantics*, 2018, pp. 1–8.
- [7] V. Kumar, Y. Hua, G. Ramakrishnan, G. Qi, L. Gao, Y.-F. Li, Difficulty-controllable multi-hop question generation from knowledge graphs, in: *International Semantic Web Conference*, Springer, 2019, pp. 382–398.
- [8] S. Elkins, E. Kochmar, J. C. K. Cheung, I. Serban, How useful are educational questions generated by large language models?, *arXiv preprint arXiv:2304.06638* (2023).
- [9] L. O. Wilson, Anderson and krathwohl bloom’s taxonomy revised understanding the new version of bloom’s taxonomy, *The Second Principle 1* (2016) 1–8.
- [10] S. Ullrich, M. Geierhos, Using bloom’s taxonomy to classify question complexity, in: *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, 2021, pp. 285–289.
- [11] M. O. Gani, R. K. Ayyasamy, A. Sangodiah, Y. T. Fui, Bloom’s taxonomy-based exam question classification: The outcome of cnn and optimal pre-trained word embedding technique, *Education and Information Technologies* 28 (2023) 15893–15914. doi:10.1007/s10639-023-11842-1.
- [12] N. Duong-Trung, X. Wang, M. Kravčík, Bloomllm: Large language models based question generation combining supervised fine-tuning and bloom’s taxonomy, in: *Lecture Notes in Computer Science*, volume 15160 LNCS, 2024, pp. 93–98. doi:10.1007/978-3-031-72312-4_11.
- [13] N. Scaria, S. Dharani Chenna, D. Subramani, Automated educational question generation at different bloom’s skill levels using large language models: Strategies and evaluation, in: *Lecture Notes in Computer Science*, volume 14830 LNAI, 2024, pp. 165–179. doi:10.1007/978-3-031-64299-9_12.
- [14] S. Maity, A. Deroy, S. Sarkar, Can large language models meet the challenge of generating school-level questions?, *Computers and Education: Artificial Intelligence* 8 (2025). doi:10.1016/j.caeai.2025.100370.
- [15] S. Bulathwela, M. Pérez-Ortiz, C. Holloway, M. Cukurova, J. Shawe-Taylor, Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools, *Sustainability* 16 (2024).
- [16] C. A. Eden, O. N. Chisom, I. S. Adeniyi, Integrating ai in education: Opportunities, challenges, and ethical considerations, *Magna Scientia Advanced Research and Reviews* 10 (2024) 006–013. doi:10.30574/msarr.2024.10.2.0039.
- [17] F. Fawzi, S. Amini, S. Bulathwela, Small generative language models for educational question generation, in: *In Proc of NeurIPS Workshop on Generative AI for Education (GAIED@NeurIPS)*, 2023.
- [18] S. Bulathwela, D. Van Niekerk, J. Shipton, M. Perez-Ortiz, B. Rosman, J. Shawe-Taylor, Truereason: An exemplar personalised learning system integrating reasoning with foundational models, *arXiv preprint arXiv:2502.10411* (2025).
- [19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* (2019).
- [20] 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 (2020) 1877–1901.
- [21] S. Elkins, E. Kochmar, J. C. K. Cheung, I. Serban, How teachers can use large language models and bloom’s taxonomy to create educational quizzes, in: *AAAI Conference on Artificial Intelligence*, 2024.
- [22] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, *Journal of Machine Learning Research* 25

- (2024) 1–53.
- [23] A. A. Yahya, Z. Toukal, A. Osman, Bloom’s taxonomy-based classification for item bank questions using support vector machines, in: *Studies in Computational Intelligence*, volume 431, 2012. doi:10.1007/978-3-642-30732-4_17.
 - [24] A. Sangodiah, R. Ahmad, W. F. W. Ahmad, Taxonomy based features in question classification using support vector machine, *Journal of Theoretical and Applied Information Technology* 95 (2017).
 - [25] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, *IBM Research Report* 22176 (2001).
 - [26] S. Banerjee, A. Lavie, Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments, *Proceedings of ACL-WMT* (2004) 65–72.
 - [27] M. Perez-Ortiz, E. Novak, S. Bulathwela, J. Shawe-Taylor, An ai-based learning companion promoting lifelong learning opportunities for all, *arXiv preprint arXiv:2112.01242* (2021).

A. Performance of the Bloom’s Taxonomy Level Classifier

The overall predictive accuracy on the one/few-shot LLMs vs. sLM is presented in figure 1

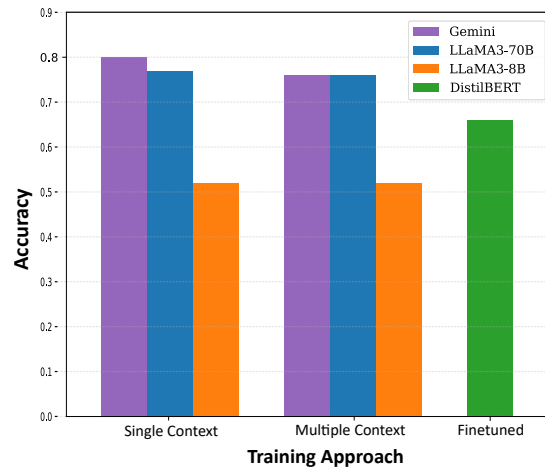


Figure 1: Overall accuracy on the held-out CGD test subset. Top classifiers achieve ~80% accuracy.