BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom's Taxonomy

Nghia Duong-Trung[1,2(✉)], Xia Wang[1], and Miloš Kravčík[1]

[1] Educational Technology Laboratory, German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91C, 10559 Berlin, Germany
{nghia_trung.duong,xia.wang,milos.kravcik}@dfki.de
[2] IU International University of Applied Sciences, Berlin Campus, Frankfurter Allee 73A, 10247 Berlin, Germany
nghia.duong-trung@iu.org

Abstract. Adaptive assessment is challenging, and considering various competence levels and their relations makes it even more complex. Nevertheless, recent developments in artificial intelligence (AI) provide new means of addressing these relevant issues. In this paper, we introduce BloomLLM, a novel adaptation of Large Language Models (LLMs) specifically designed to enhance the generation of educational content in alignment with Bloom's Revised Taxonomy. BloomLLM performs well across all levels of competencies by providing meaningful, semantically connected questions. It is achieved by addressing the challenges of foundational LLMs, such as lack of semantic interdependence of levels and increased hallucination, which often result in unrealistic and impractical questions. BloomLLM, fine-tuned on ChatGPT-3.5-turbo, was developed by fine-tuning 1026 questions spanning 29 topics in two master courses during the winter semester 2023. The model's performance, outpacing ChatGPT-4, even with varied prompting strategies, marks a significant advancement in applying generative AI in education. We have publicly made the BloomLLM codes and training datasets available to promote transparency and reproducibility.

Keywords: Bloom's Revised Taxonomy · BloomLLM · ChatGPT · Supervised Fine-Tuning · LLMs

1 Introduction

Recent trends show a significant influx of LLMs into educational settings, spurring academic interest in the synergy between humans and AI. This collaboration is hypothesized to foster more effective learning environments than those created by either entity in isolation [3]. Educators are therefore bracing for a paradigm shift in educational dynamics, particularly in the roles of teachers,

R. Ferreira Mello et al. (Eds.): EC-TEL 2024, LNCS 15160, pp. 93–98, 2024.
https://doi.org/10.1007/978-3-031-72312-4_11

as AI begins to assume functions traditionally reserved for human educators, enhancing learning experiences in ways that surpass both conventional methods and existing LLM capabilities [4,5]. However, the precise nature and extent of ChatGPT's impact on learning, both positive and negative, remain subjects of ongoing debate and are yet to be thoroughly investigated [5,8].

Our research engages with these novel educational paradigms, particularly in the context of learning enhancement through innovative content and materials. Utilizing the Revised Bloom's Taxonomy (2001) [1], which categorizes educational objectives into cognitive and affective domains, our study aligns with the six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. This taxonomy is a foundational framework to deepen students' comprehension of subject matter systematically. Implementing this model necessitates meticulous planning to scaffold learning materials across different cognitive stages appropriately [9]. A noted challenge in this regard is the generation of interdependent assessment tasks tailored to individual student proficiency levels.

This paper explores three critical research questions (RQ) concerning the creation of evaluation tasks that are in harmony with the Revised Bloom's Taxonomy: **RQ1**: How effective is ChatGPT in producing tasks (questions) that align with Bloom's taxonomy, and is it practical for educators to use? **RQ2**: What methods can educators use to guide ChatGPT towards achieving better outcomes for specific educational goals, especially when initial attempts are unsatisfactory? **RQ3**: What are the approaches for integrating ChatGPT or comparable generative AI technologies within educational settings to assist teachers and students comprehensively? To the best of our knowledge, our research is pioneering in integrating Bloom's Taxonomy with LLMs foundation models for personalized and customizable question generation in higher education.

2 Backgrounds and Related Work

Incorporating Bloom's Taxonomy into AI in educational contexts represents a multifaceted strategy. This approach augments the learning experience and streamlines lesson preparation for educators. Contemporary research has explored the utilization of LLMs and ChatGPT to formulate learning assessment aligned with Bloom's Taxonomy. For instance, Kwan's investigation employed ChatGPT to generate assessment scripts, including marking schemes and solutions in Probability and Engineering Statistics, guided by Bloom's framework [6]. However, it was noted that educational professionals need to further scrutinize and refine these outputs to ensure appropriate differentiation in question difficulty levels. Herrmann-Werner and colleagues investigated the effectiveness of ChatGPT in addressing questions formulated based on Bloom's Taxonomy, applying a variety of prompts [2]. This examination assessed ChatGPT's performance when presented with differing input types. In a parallel, Morjaria *et al.* probed into ChatGPT's proficiency in handling brief evaluative questions within the context of an undergraduate medical course [7]. These studies collectively

represent initial attempts to integrate Bloom's Taxonomy with ChatGPT's functionalities. Yet, these explorations are in their nascent stages of employing ChatGPT to generate questions aligned with Bloom's taxonomy, without specifically tailoring the output for a particular audience or academic program.

3 BloomLLM

3.1 The Current Challenges of ChatGPT in Education

Central to Bloom's taxonomy is the progression from lower-order to higher-order thinking. Significantly, the Applying level, the third in this hierarchy, serves as a pivotal juncture, bridging the foundational stages of Remembering and Understanding with the more advanced stages of Analyzing, Evaluating, and Creating. In this context, assessing the outputs of ChatGPT-4, as presented in Tables 1 and 2 in our GitHub repository https://github.com/duongtrung/BloomLLM, comparing to our students' competence and curriculum, we have several observations: (i) ChatGPT demonstrates proficiency in generating questions that align with the Remembering and Understanding levels of Bloom's taxonomy; (ii) A notable limitation of ChatGPT is its lack of interdependence among cognitive levels. It fails to establish semantic connections between these levels, which are essential for facilitating a cohesive educational journey from lower to higher cognitive processes; (iii) ChatGPT's performance at the Applying level, a crucial intermediary stage, is found to be overly simplistic, and (iv) ChatGPT generated unrealistic and undoable questions in the Evaluating and Creating levels, even for a Ph.D. Despite exploring various topics and employing complex prompts, the authors find that ChatGPT-4 does not consistently produce the anticipated responses. Consequently, we address our first research question **RQ1**: Is ChatGPT sufficiently capable to be employed by educators? The answer, as our findings suggest, is negative. Readers are encouraged to briefly review the analysis in Sect. 3.4. Fundamentally, ChatGPT generates token sequences based on word probability from its training, disconnected from the specifics of educational institutions, programs, or desired learning outcomes. This raises the question of how to infuse domain-specific knowledge to influence word probability. Our attention now shifts towards offering solutions, thereby addressing our **RQ2**.

3.2 Supervised Fine-Tuning

We denote the tokens for the i^{th} prompt by $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \dots]$ and the tokens in the human-written response by $\boldsymbol{y}_i = [y_{i,1}, y_{i,2}, \dots, y_{i,T_i}]$, the loss function is $L[\theta] = -\sum_{i=1}^{I} \sum_{j=1}^{T_j} \log[P(y_{i,t+1}|\boldsymbol{x}_i, y_{i,1\dots t}, \theta)]$, where θ is the model parameters. The model aims to maximize the probability of the response tokens and generalize the unseen prompt $\boldsymbol{x}_u$.

SFT shows remarkable effectiveness when the outcomes defined by humans, represented as $\boldsymbol{y}$, are clearly specified. A visual representation of SFT is shown in Fig. 1. In this depiction, tokens correspond to a set of subjects $\boldsymbol{x}$, and the

goals align with the answers to Bloom's taxonomy questions crafted by experts, denoted as $\boldsymbol{y}$. The element $\boldsymbol{y}$ encompasses six levels within Bloom's taxonomy, each featuring a series of questions. Experts are tasked with creating tailored and individualized questions based on Bloom's taxonomy, considering factors such as the curriculum, undergraduate and graduate programs, desired learning outcomes, and student capabilities.

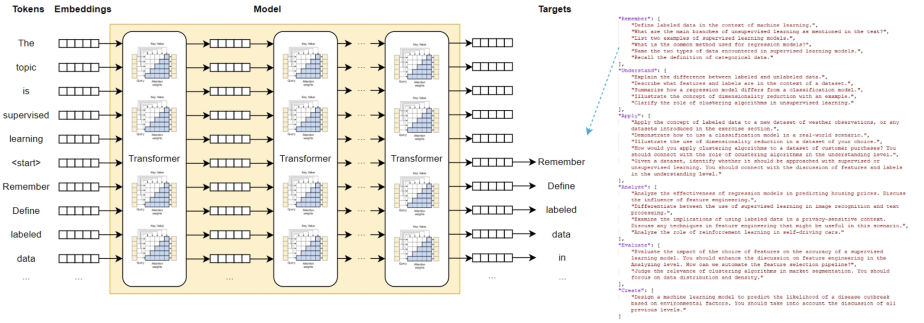

Fig. 1. The pre-trained LLMs model is fine-tuned using a pairs of topic $\boldsymbol{x}$ and expert-written responses to Bloom's questions or completion $\boldsymbol{y}$.

At our university, we have historically developed learning control questions (assessment tasks) in all courses. Lecturers and professors can manually develop questions to help students understand the lessons better. Seeking to streamline this process, we have turned to generative AI to augment teaching activities and facilitate scalable deployment. To uphold content copyright and maintain the confidentiality of specific datasets, tables, and figures from our textbooks, we have crafted 1026 questions covering 29 topics across two master courses: *Introduction to Machine Learning*, and *Neural Networks and Deep Learning*, during the winter semester 2023. They resulted in 47 samples, e.g., 47 pairs of $\boldsymbol{x}$ and $\boldsymbol{y}$. These questions are manually annotated by lecturers who gave those lectures in the academic years 2022 and 2023. The tasks satisfy Bloom's taxonomy, encompass coding, mathematical exercises, and semantic links between Bloom levels, and are targeted at the appropriate students. The sets were uploaded to OpenAI's servers for fine-tuning ChatGPT-3.5-Turbo-1106. Here, we address **RQ2**.

Following the fine-tuning process, we deployed BloomLLM and made it available as a service on the OpenAI Platform for Enterprise, benefiting over 4000 university staff, professors, and teachers. The platform provides convenient endpoints for any GPT services developed in-house. Given the breadth of our university's academic offerings, which include approximately 200 Bachelor's, Master's, and MBA degree programs, it is imperative that any deployed AI solutions be easily accessible to faculty members, regardless of their IT proficiency. It enables them to concentrate on educational content rather than technical intricacies. In

this context, we address **RQ3** by utilizing the proprietary platform for rapid and efficient deployment at scale.

3.3 Experimental Results

The dataset was partitioned into training and test subsets in an 80–20 ratio. This partitioning was executed three times to mitigate potential biases in the dataset-splitting process. The outcomes of these tasks have been documented, including average values and standard deviations. The train accuracy, test accuracy, train loss, test loss, total tokens, and total cost are 0.8011 ± 0.0159, 0.6844 ± 0.0153, 0.6708 ± 0.0648, 0.9534 ± 0.0504, 330506, and $\sim$23.79 USD, respectively.

3.4 Evaluation

The study scrutinized how 46 master's students from three distinct classes perceived and valued a series of questionnaire-based assessments to gauge their receptiveness. These questionnaires, tailored to four topics, explored the students' eagerness to engage with the questions, evaluating their relevance and utility concerning their academic programs, learning outcomes, and curriculum integration. Among the subjects, *Machine Learning Introduction* was a part of BloomLLM's original training dataset, whereas the inclusion of *Introduction to Data Science*, *Pattern Recognition*, and *Time Series Forecasting* served to test BloomLLM's extrapolative capabilities, as these topics were not covered in its initial training data. BloomLLM and ChatGPT-4 were assigned to formulate question sets for each topic. These sets comprised evaluation tasks spanning six levels of Bloom's taxonomy, with three questions per level. Students were allowed to endorse both question sets if they found the assessments meaningful or, alternatively, to abstain from selecting any, thus indicating a neutral or indecisive

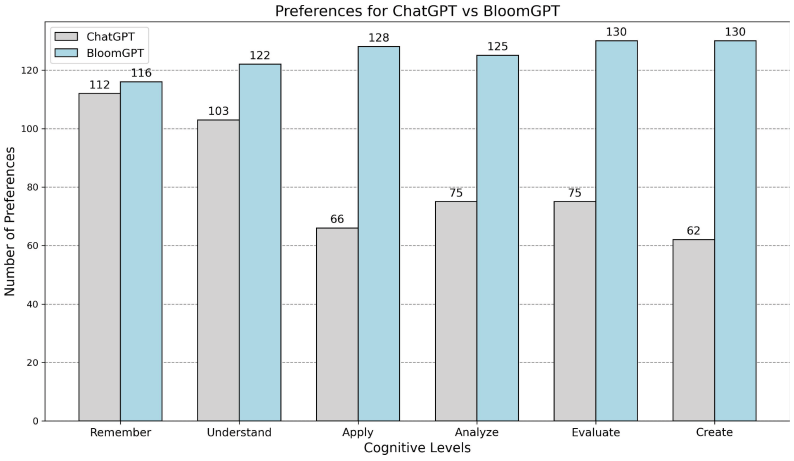

Fig. 2. Preferences for ChatGPT vs BloomLLM across cognitive levels.

stance. The essence of the research, focusing on the student's preferences and perceptions on both a general and a nuanced, level-specific basis, was illustrated in columnar plots depicted within Fig. 2.

4 Conclusion

In this paper, we have addressed three research questions regarding the effectiveness of ChatGPT in generating educational content aligned with Bloom's Taxonomy. Our proposed BloomLLM emerges as a solution to the initial non-effectiveness of standard LLMs in the academic domain. It not only excels in generating semantically rich and interconnected questions across the taxonomy's spectrum but also sets a new standard in educational AI, outperforming ChatGPT-4. We release BloomLLM to all educators and master students at our university and commit to maintaining and updating it with new courses and questions. By making the codes and datasets publicly available, we have laid a foundation for widespread academic use, as evidenced by the rapid adoption within our university community.

Acknowledgement. The authors kindly appreciate the support of CATALPA, FernUniversität in Hagen through the"AI.EDU Research Lab 2.0" Project.

References

1. Anderson, L.W., et al.: A taxonomy for teaching, learning, and assessment (2001)
2. Herrmann Werner, A., et al.: Assessing chatGPT's mastery of bloom's taxonomy using psychosomatic medicine exam questions. medRxiv **8** (2023)
3. Jeon, J., Lee, S.: Large language models in education: a focus on the complementary relationship between human teachers and chatGPt. Educ. Inf. Technol. **28**(12), 15873–15892 (2023)
4. Kaplan-Rakowski, R., Grotewold, K., Hartwick, P., Papin, K.: Generative AI and teachers' perspectives on its implementation in education. J. Interact. Learn. Res. **34**(2), 313–338 (2023)
5. Kasneci, E., et al.: ChatGPT for good? on opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
6. Kwan, C.C.L.: Exploring chatGPT-generated assessment scripts of probability and engineering statistics from bloom's taxonomy. In: International Conference on Technology in Education, pp. 275–286. Springer (2023). https://doi.org/10.1007/978-981-99-8255-4_24
7. Morjaria, L., et al.: Examining the threat of chatGPT to the validity of short answer assessments in an undergraduate medical program. J. Med. Educ. Curric. Dev. **10**, 23821205231204176 (2023)
8. Tlili, A., et al.: What if the devil is my guardian angel: chatgpt as a case study of using chatbots in education. Smart Learn. Environ. **10**(1), 15 (2023)
9. Wang, X., et al.: IFSE-personalized quiz generator and intelligent knowledge recommendation. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC), pp. 201–208. IEEE (2022)