Harnessing LLMs in Curricular Design: Using GPT-4 to Support Authoring of Learning Objectives

Pragnya Sridhar[1,*,†], Aidan Doyle[1,*,†], Arav Agarwal[2], Christopher Bogart[2], Jaromir Savelka[2] and Majd Sakr[2]

[1]*Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA*

[2]*Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract

We evaluated the capability of a generative pre-trained transformer (GPT-4) to automatically generate high-quality learning objectives (LOs) in the context of a practically oriented university course on Artificial Intelligence. Discussions of opportunities (e.g., content generation, explanation) and risks (e.g., cheating) of this emerging technology in education have intensified, but to date there has not been a study of the models' capabilities in supporting the course design and authoring of LOs. LOs articulate the knowledge and skills learners are intended to acquire by engaging with a course. To be effective, LOs must focus on what students are intended to achieve, focus on specific cognitive processes, and be measurable. Thus, authoring high-quality LOs is a challenging and time consuming (i.e., expensive) effort. We evaluated 127 LOs that were automatically generated based on a carefully crafted prompt (detailed guidelines on high-quality LOs authoring) submitted to GPT-4 for conceptual modules and projects of an AI Practitioner course. We analyzed the generated LOs if they follow certain best practices such as beginning with action verbs from Bloom's taxonomy in regards to the level of sophistication intended. Our analysis showed that the generated LOs are sensible, properly expressed (e.g., starting with an action verb), and that they largely operate at the appropriate level of Bloom's taxonomy, respecting the different nature of the conceptual modules (lower levels) and projects (higher levels). Our results can be leveraged by instructors and curricular designers wishing to take advantage of the state-of-the-art generative models to support their curricular and course design efforts.

Keywords

GPT-4, Large Language Models, LLMs, Learning Objectives, Automatic Generation, Curricular Development, Course Design Automation, Automated Content Generation

1. Introduction

Learning objectives (LOs) are the blueprints against which course content is designed. They provide instructors with a framework for content curation, instructional and assessment strategies, as well as enable learners to reflect on and plan their own learning of a course's knowledge and skills. Alignment between LOs, instructional strategies and assessments is a necessary

Empowering Education with LLMs – the Next-Gen Interface and Content Generation, July 7, 2023, Tokyo, Japan

[*]Corresponding author.

[†]These authors contributed equally.

pragnyas@andrew.cmu.edu (P. Sridhar); adoyle@andrew.cmu.edu (A. Doyle); arava@andrew.cmu.edu (A. Agarwal); cbogart@andrew.cmu.edu (C. Bogart); jsavelka@cs.cmu.edu (J. Savelka); msakr@cs.cmu.edu (M. Sakr)

0000-0002-9421-8566 (C. Bogart); 0000-0002-3674-5456 (J. Savelka)

CEUR Workshop Proceedings (CEUR-WS.org)

prerequisite for internally consistent learning experience. When the three components are misaligned, learners may feel that tests do not assess what was covered in class, or instructors may notice that students earn a passing grade without mastering the material at the desired level. Hence, poor quality or missing LOs have a negative impact on the learning experience.[1]

Creating effective LOs can be challenging and time-consuming for instructors, requiring substantial knowledge and experience in instructional design. Bloom [1, 2, 3] proposed a taxonomy organizing LOs into six levels (remember, understand, apply, analyze, evaluate, and create). This taxonomy helps educators articulate LOs that focus on concrete actions and behaviors, and target distinct levels of cognitive processes. For LOs to guide the selection of assessments, they must be measurable, i.e., it should be possible to evaluate whether learners attained the intended objective. Because of the complexity involved in authoring high-quality LOs, instructors often forego the task in lieu of more pressing duties such as authoring the course content or teaching.

Furthermore, instructors may have general notions of the learning objectives they want students to accomplish by the course's end. However, these notions may not meet the criteria of being well-defined and measurable learning objectives that focus on what a learner will achieve. To address this issue, our aim is to develop an approach that generates high-quality learning objectives and streamlines the objective-setting process. In our initial experiment, we investigate the potential of utilizing Language Models (LLMs) to generate effective learning objectives. This exploration serves as a stepping stone for future solutions that involve refining and augmenting the learning objectives initially provided by instructors.

Large Language Models (LLMs) are sophisticated AI models pre-trained on extensive textual data, that can generate human-level quality text. With appropriate prompting, an LLM might create high-quality LOs, alleviating the burden on the instructors. In this study we explore the potential of a state-of-the-art LLM (GPT-4) to support this task. We hypothesize that a well-prompted LLM can more efficiently produce candidate LOs, saving instructors' time.

To investigate the capability of GPT-4 to efficiently generate quality LOs in the context of a software development course on the practical integration of AI into applications, we analyzed the following research questions:

RQ1: Are the generated LOs sensible, i.e., clear grammatically correct statements, addressing the relevant topic?

RQ2: Do the LOs start with an appropriate action verb describing measurable behavior?

RQ3: Do the conceptual module and project-related LOs target cognitive processes at the appropriate levels of Bloom's taxonomy?

To our best knowledge, this is the first study that proposes and evaluates automatic generation of LOs to drive the course design process, as opposed to generating LOs from an already existing course materials [4].

[1] Eberly Center: Design & Teach a Course. Available at: https://www.cmu.edu/teaching/designteach/design/learningobjectives.html [Accessed 2023-05-10]

2. Related Work

Several works posit that LLMs could be an asset in pedagogy in several ways including Assessment Generation, Personalized Feedback System, generating lesson plans and asking questions about the best ways to teach a subject [5, 6, 7]. While much work has already been done towards the applications of LLMs and course-content generation, LLMs have yet to be shown to generate course-guiding LOs. Thus, our focus on generating course-guiding LOs by leveraging the abilities of LLMs.

Tran et al. used IBM Watson to automatically generate LOs based on key-phrases extracted from course material documents paired with action verbs from Bloom's Taxonomy [4]. In this work, we generate LOs using state-of-the-art LLM (GPT-4). While the system proposed in [4] employs a similar definition of LOs to ours, the intended use is different. We generate LOs prior to course materials to potentially guide the content generation process. The system from [4] assumes an existing collection of course material, and generate LOs from the collection.

Other work that harnesses LLMs to generate course material includes Multiple-Choice Question (MCQ) generation, such as the Question-Answer-Distractor pipeline in [8]. Lu et al. utilized LLMs to effectively generate reading quizzes, confirming the effectiveness of the system in manual evaluation by 11 instructors across 7 different universities [9].

Adams comments on the direct application of Bloom's Taxonomy to developing LOs [3]. By including action verbs associated with different levels of the taxonomy, educators are encouraged to think in terms of what students should be able to do at the end of the course. Additionally, the authors of [4] generated the LOs by training a multi-layer perceptron (MLP) to learn which Bloom's Taxonomy action verb would best fit with each of the key phrases extracted from the collection of course material documents[4]. In this work, the evaluation heavily focuses on verifying that the generated learning objectives begin with an action verb from an appropriate level of Bloom's Taxonomy.

In computing education context, LLMs have been shown to be highly effective at generating code and explanations of the code for entry-level programmers [10]. Such explanations have even been observed to out-class student explanations of the same code [11]. Denny et al. discovered that well-structured prompts could yield correct solutions to many programming exercises [12]; observations later confirmed by Savelka et al. [13, 14]. Piccolo et al. demonstrated that LLMs can perform most entry-level programming tasks in the context of introductory bioinformatics course [15]. However, Savelka et al. [16] and Wermelinger [17] point to some limitations of LLMs in handling assessments from introductory programming classes. Phung et al. introduced a system that harnessed LLMs to provide precision feedback on syntax errors in students' code [18]. Such feedback explanations went far beyond describing the code line-by-line. Sarsa et al. found such explanations were particularly valuable for student learning [10]. MacNeil et al. demonstrated that explanations of generated code can be offered at multiple different levels of abstraction [19, 20]. In the near future, it is reasonable to expect LLMs to facilitate teacher-student exchanges similar to those that only occur in a classroom, and are invaluable to student learning [21].

3. Experiments

3.1. Model

GPT has gained significant popularity due to its remarkable advancements in understanding and generating natural language text. It has demonstrated superior performance across various domains, including code generation [22], software engineering [23], solving AI tasks [24], and data augmentation [25], showcasing its domain-invariant capabilities. In the context of education, a study conducted by Malinka et al. [7] specifically investigated the impact of ChatGPT, a variant of the GPT model, on higher education with a focus on computer programming subjects. The authors provided evidence highlighting the effectiveness of ChatGPT in managing programming assignments, exams, and homework tasks.

Building on the success of its predecessors, GPT-4 represents a significant leap forward in language modeling technology [26]. Thus in our pursuit of generating course syllabus, we use the GPT-4 model (gpt-4). As of the writing of this paper, GPT-4 is by far the most advanced model released by OpenAI. The model is focused on dialog between a user and a system (i.e., an assistant).

We set the `temperature` of the model to 0.7, which is the default. The higher the `temperature` the more creative the output but it can also be less factual. As the temperature approaches 0.0, the model becomes deterministic and can be repetitive. We set `max_tokens` to 2,000 tokens (a token roughly corresponds to a word). This parameter controls the maximum length of the completion (i.e., the output). Note that GPT-4 has an overall token length limit of 8,192 tokens, comprising both the prompt and the completion.[2] We set `top_p` to 1 (default). This parameter is related to `temperature` and also influences creativeness of the output. We set `frequency_penalty` to 0, which allows repetition by ensuring no penalty is applied to repetitions. Finally, we set `presence_penalty` to 0, ensuring no penalty is applied to tokens appearing multiple times in the output.

3.2. Experimental Design

To generate LOs, we utilize the system prompt illustrated in Figure 2. The system prompt guides the GPT-4 model towards the desired behavior. The prompt contains brief guidelines on how LOs should be structured and what properties are desirable. These guidelines were informed by various university materials on course design.[3] The guidelines instruct the system to generate LOs for conceptual modules as well as projects. The LOs should start with an action verb describing the behavior, state the conditions under which it is to be performed, and the degree of mastery the learners should attend. The prompt also provides many example LOs. These can be related to conceptual modules that focus on the two lower levels of Bloom's taxonomy, e.g.:

> Define DevOps from organizational, cultural and technical perspectives.

[2] There is also a variant of the model that supports up to 32,768 tokens.

[3] Eberly Center: Design & Teach a Course. Available at: https://www.cmu.edu/teaching/designteach/design/learningobjectives.html [Accessed 2023-05-10]; Center for Excellence in Teaching and Learning at UCONN: Developing LOs. Available at: https://cetl.uconn.edu/resources/design-your-course/developing-learning-objectives/ [Accessed 2023-05-10]

```
You are a curricular development expert system focused on authoring LOs. Learning
objectives are brief, clear statements that describe the desired learning outcomes of instruction.
[601 characters ...] LOs should use action verbs. LOs should be
measurable.

A well-constructed learning objective contains three parts [392 characters ...]

1. BEHAVIOR
The behavior is the real work to be accomplished by the student specified by an action verb that
connotes observable and measurable behaviors. [2,497 characters ...]

2. CONDITIONS
This is a statement that describes the exact conditions under which the defined behavior is to be
performed. [117 characters ...]

3. DEGREE
This is a statement that specifies how well the student must perform the behavior [171 characters ...]

Conceptual LOs are focused on students' knowledge and understanding (i.e., the first
two levels of Bloom's taxonomy).

[18 example LOs (1,540 characters)  ...]

Project LOs are focused on students' skills and behaviors (i.e., the higher levels of
Bloom's taxonomy).

[12 example LOs (1,261 characters)  ...]

Here are some criteria to satisfy in order to create an effective learning objective:
1. LOs should be student-centered. [114 characters ...]
2. LOs should focus on specific cognitive processes. [530 characters ...]
3. LOs should use action verbs.
4. LOs should be measurable [105 characters ...]

The user will provide you with the name of the course, brief description of the course goals, the name
of the module, and the type of the LOs to be developed. Based on these you respond with
a list of well-designed effective LOs (5-10 items).
```

Figure 2: The system prompt. The figure shows the essential elements of the system prompt guiding the model towards generating high-quality LOs. The grey annotations are used as replacement for extensive parts of the prompt that could not be fit into the figure.

The LOs from projects focus on behaviors described by action verbs from higher levels of Bloom's taxonomy, e.g.:

> Design and implement Continuous Integration and Continuous Delivery for a Node.JS application.

The prompt emphasizes the difference between the conceptual module and project related LOs. Hence, given the same topic the LOs for the conceptual module are expected to differ

```
COURSE NAME: AI Practitioner
COURSE GOALS: In this course, learners gain hands-on experience solving real-world problems by
completing projects focused on developing AI/ML-enabled systems. It is our goal that students will
develop the skills needed to become sophisticated developers of AI/ML-based systems. Specifically,
students are exposed to real-world data and scenarios to learn how to:
- Integrated different types of AI/ML systems into their applications, recognize their capabilities and
limitations.
- Explain the effects of data quality, quantity, and representativeness on the performance of AI/ML
systems.
- Inspect, validate, and critically assess outputs of AI/ML systems.
- Use AI/ML components via cloud APIs or locally run libraries from different areas, such as language
technologies, or computer vision, including state-of-the art generative models.
- Discuss the advantages and disadvantages of different computing devices and environments for
deployment of Artificial Intelligence systems.
Through this process, we aspire for our students to become sophisticated, independent, and resilient
problem solvers who are able to overcome challenges and learn.
MODULE NAME: {{module}}
LOs TYPE: {{module_type}}
EXPECTED OUTPUT:
1. Text of LO1.
2. Text of LO2.
...
```

Figure 3: The User Message. The figure shows the user message template specifying the context for the LOs to be generated. The grey orange tokens are replaced with module specific information.

substantially from those generated for the project in terms of their focus on different levels of Bloom's taxonomy.

The context of the particular course being designed, i.e., AI Practitioner, is provided to GPT-4 via the user message. The full template of the user message used in this work is shown in Figure 3. It provides the name of the course, brief description of the high-level course goals, placeholders for a module name (e.g., "Generative Models" or "AI/ML in the Cloud") and module type (i.e., either a "conceptual module" or a "project"). To generate LOs for each conceptual module and each project, a separate message is used with the placeholders filled in accordingly. The dynamically constructed prompts, i.e., the system prompt and the user message, are submitted individually to OpenAI's GPT-4 API using the `openai` Python library.[4]

We extracted the generated LOs from the GPT-4 responses and analyzed them to answer the three research questions. To answer RQ2, we used a simple regular expression to extract the action verb from each learning objective. To answer RQ3, we evaluated the generated LOs using both automatic methods and human annotation. We presented the 127 of the LOs to 3 graduate computer science students and asked them to classify them into the individual levels of Bloom's Taxonomy. Out of these, 101 LOs were annotated by all three of the annotators. We also automatically classified the generated LOs into the individual levels of Bloom's taxonomy using the approach mentioned in [27]. They trained a binary classifier for each Bloom's Taxonomy category using a dataset of 21,380 LOs from 5,558 university courses. We used the same models

[4] GitHub: OpenAI Python Library. Available at: https://github.com/openai/openai-python [Accessed 2023-05-10]

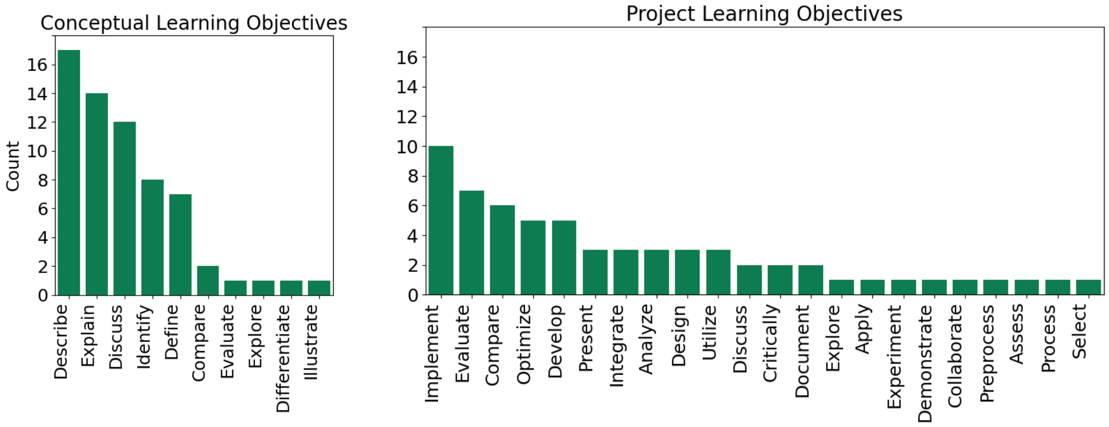

Figure 4: Distribution of action verbs across module types.

to predict the Bloom's Taxonomy level of the generated LOs.

3.3. Results

Figure 4 shows that LOs for conceptual modules heavily employ a small number of action verbs such as "describe," "discuss," "explain," "identify" and "define". This is expected since these verbs are geared toward conceptual learning. The LOs for projects use more diverse set of action verbs. The examples include "implement," "optimize," "develop," or "utilize." These verbs also appear sensible as they focus on activities and skills as opposed to conceptual knowledge.

Figure 5 reports the distribution of LOs across the conceptual modules and projects categorized in terms of Bloom's Taxonomy categories, using the classifier proposed in [27]. The LOs generated for conceptual modules predominantly fall under the "Understand" level, while the LOs generated for projects are spread across the "Apply," "Analyze," "Evaluate," and "Create" levels. This is expected. Figure 6 shows similar distribution where instead of the BERT-based classifier we use levels of Bloom's taxonomy assigned by each of our human annotators. Note that the counts in this figure are normalized, as each LO was annotated three times during human annotation. The distribution differs a bit but we can still observe the LOs for the conceptual modules mostly employing the action verbs from the lower levels of Bloom's taxonomy, whereas the LOs for projects use action verbs from the higher levels.

4. Discussion

4.1. RQ1: Are the generated LOs sensible?

Overall, the LOs are largely sensible. They describe key sub-concepts related to the relevant topic, and mostly focus on one or two separate cognitive processes, e.g.:

> Explain the key concepts and techniques used in computer vision, such as image processing, feature extraction, and object recognition.

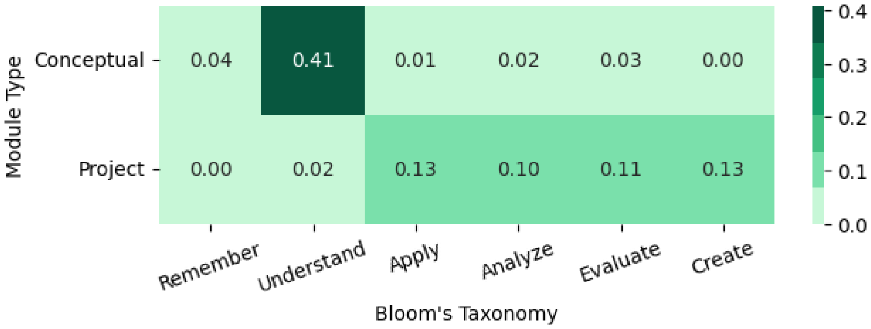

Figure 5: Bloom's Taxonomy Categorization of Generated LOs using BERT

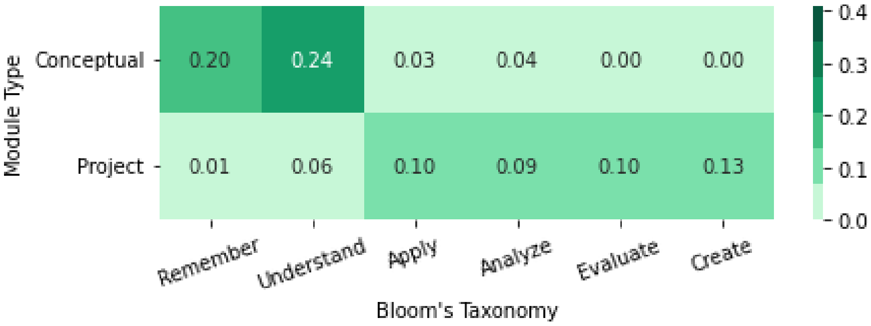

Figure 6: Bloom's Taxonomy Categorization of Generated LOs based on Human Annotation

GPT-4 generates LOs with action verbs such as "describe," "explain," and "discuss" for conceptual modules, and "implement," "evaluate" and "develop" for project-based materials. While the LOs are sensible they sometimes lack specific focus. For example, one such generated LO was "Implement a basic AI/ML model using Python libraries to solve a simple classification or regression problem." While this is measurable, the term "Python libraries" is too broad, and would benefit from being more focused (e.g., "scikit-learn"). This issue may be solvable with further prompt tuning.

4.2. RQ2: Do the LOs start with an appropriate action verb?

All the LOs start with action verbs. The distribution of action verbs across LOs for conceptual modules and projects is as expected. Action verbs such as "describe" and "explain" should be associated with conceptual materials often focused on declarative knowledge. Whereas action verbs such as "implement" and "develop" should be associated with projects geared towards procedural knowledge.

We found that some generated LOs (Figure 4) employed action verbs that were not included in the extensive list of example verbs provided to the model via the prompt. Specifically, these were "optimize," "preprocess," "explore," "document," "implement," "utilize," and "process". Of these, only “utilize” and “implement” appeared in the examples mentioned in the prompt. Of the 127, there were 26 LOs starting with action verbs not included in the list of examples: 25 of these were for projects and one was for a conceptual module. Moreover, there were 13 LOs with action verbs that have that were not provided anywhere in the prompt.

Furthermore, there are 11 generated LOs that utilize multiple action verbs, e.g., “Evaluate the performance of computer vision models using appropriate metrics and develop strategies to improve their accuracy and reliability.” These could have been separate LOs (e.g., “Evaluate the performance of computer vision models using appropriate metrics.” and “Develop strategies to improve the accuracy and reliability of computer vision models”).

4.3. RQ3: Do the LOs target cognitive processes at the appropriate levels of Bloom’s taxonomy?

The results of applying BERT classifier as described in Section 3.2 show that the GPT-4 model generates LOs that largely operate on the expected levels of Bloom’s Taxonomy (Figure 5). Conceptual modules are more focused on declarative knowledge, and the LOs mostly employ action verbs from the lower two levels of Bloom’s taxonomy (Remember and Understand). Projects are focused on procedural knowledge, and the LOs mostly use action verbs from the higher four levels (Apply, Analyze, Evaluate and Create). Based on the BERT classifier, it appears that the generated LOs are geared towards the appropriate types of cognitive processes.

The classifier proposed in [27] may have some limitations. We noticed that two of the generated LOs were not assigned to any of the Bloom’s Taxonomy levels, while five LOs were assigned multiple categories. Note that these issues pertain to a relatively small proportion of the generated LOs.

We presented the generated LOs to human annotators to validate the BERT-based classification. As seen in Figure 5, most LOs generated for conceptual modules were classified as targeting the Understand and Remember levels of Bloom’s taxonomy, and most project LOs were classified as *not* Understand or Remember. This same distribution can be observed in 6, although with a higher frequency for humans classifying conceptual LOs as ’Remember’ instead of ’Understand’. When conducting the human annotation, we used Cohen’s K to examine our inter-rater agreement. The average agreement among raters for classifying LOs into each of the six individual levels of Bloom’s Taxonomy was 0.31, corresponding to fair agreement. By mapping the BERT and human classifications to the corresponding LO categories, we were able to observe an agreement between the majority-vote annotation and the BERT classification of 0.62, demonstrating substantial agreement between the human and BERT classification of LOs into either the ’Remember’ and ’Understand’, or the ’Apply’, ’Analyze’, ’Evaluate’, and ’Create’ levels of Bloom’s Taxonomy.

5. Implications for Education Practice

Automatic generation of LOs could significantly ease the workload of educators, allowing them to focus more on teaching and student interaction. Automation could improve the quality of LOs. Furthermore, reducing the cost of authoring LOs could open up a possibility of personalized LOs for each student based on their individual strengths, weaknesses, and progress. On the other hand, the over-reliance on automated systems could potentially lead to a loss of pedagogical nuance and adaptability. Additionally, the standardization might stifle creativity and innovation in teaching methods. Therefore, integrating such systems into teaching practice should be handled with caution, ensuring they serve as a supportive tool rather than a replacement for educators' expertise.

6. Limitations

LOs drive the entire course development process. Mistakes in authoring LOs may cascade and snowball into larger issues manifesting in low-quality course content. In addition, LLMs are relatively new technology, and there may be skepticism or resistance from educators and experts regarding the reliability and validity of using LLMs to generate LOs. Addressing these concerns and gaining acceptance is essential for potential widespread adoption of the proposed approach. Additionally, there is a need for human validation of the generated LOs. While LLMs can assist in the initial generation process, human expertise and judgment are crucial to ensure the accuracy, relevance, and appropriateness of the generated LOs. Also, some instructors or institutions have suggested use of GPT to be unethical because of its training on copyrighted materials; thus; its products may not be usable in institutions that have policies preventing such use.

7. Conclusions and Future Work

This paper explored the use LLMs for generating LOs in the context of practically oriented university course on AI. Prior work demonstrated the effectiveness of LLMs in various tasks in educational context, even in generating various elements of course content. However, there is a lack of work on LLMs' capability to generate LOs. This work highlights the potential of LLMs for generating LOs to support curricular development. We evaluated the effectiveness of GPT-4 on this task. We found that the generated LOs are sensible, properly expressed (e.g., starting with an action verb), and that they largely operate at the appropriate level of Bloom's taxonomy respecting the different nature of the conceptual modules (lower levels) and projects (higher levels). These findings can be leveraged by instructors and curricular designers wishing to take advantage of the state-of-the-art generative models to support their curricular and course design efforts.

In future work, we plan to further evaluate the generated LOs, especially in terms of the LOs being measurable. This may include generating LOs for existing courses with existing human-created LOs in order to compare the two. In addition, GPT-4 could also be used for developemn of assessment strategies for the generated LOs.

References

[1] B. S. Bloom, M. D. Engelhart, E. Furst, W. H. Hill, D. R. Krathwohl, Handbook i: cognitive domain, New York: David McKay (1956).

[2] D. R. Krathwohl, A revision of bloom's taxonomy: An overview, Theory into practice 41 (2002) 212–218.

[3] N. E. Adams, Bloom's taxonomy of cognitive learning objectives., Journal of the Medical Library Association : JMLA 103 3 (2015) 152–3.

[4] K. N. Tran, J. H. Lau, D. Contractor, U. Gupta, B. Sengupta, C. J. Butler, M. Mohania, Document chunking and learning objective generation for instruction design, in: EDM, International EDM Society, 2018.

[5] J. Su, W. Yang, Unlocking the power of chatgpt: A framework for applying generative ai in education, ECNU Review of Education 0 (0) 20965311231168423. URL: https://doi.org/10.1177/20965311231168423. doi:10.1177/20965311231168423.

[6] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, 2023. URL: edarxiv.org/5er8f. doi:10.35542/osf.io/5er8f.

[7] K. Malinka, M. Perešíni, A. Firc, O. Hujňák, F. Januš, On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?, 2023. arXiv:2303.11146.

[8] R. Rodriguez-Torrealba, E. Garcia-Lopez, A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, Expert Syst. Appl. 208 (2022). URL: https://doi.org/10.1016/j.eswa.2022.118258. doi:10.1016/j.eswa.2022.118258.

[9] X. Lu, S. Fan, J. Houghton, L. Wang, X. Wang, Readingquizmaker: A human-nlp collaborative system that supports instructors to design high-quality reading quiz questions, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (2023).

[10] S. Sarsa, P. Denny, A. Hellas, J. Leinonen, Automatic generation of programming exercises and code explanations using large language models, ACM, 2022. URL: https://doi.org/10.1145%2F3501385.3543957. doi:10.1145/3501385.3543957.

[11] J. Leinonen, P. Denny, S. MacNeil, S. Sarsa, S. Bernstein, J. Kim, A. Tran, A. Hellas, Comparing code explanations created by students and large language models, 2023. arXiv:2304.03938.

[12] P. Denny, V. Kumar, N. Giacaman, Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language, 2022. arXiv:2210.15157.

[13] J. Savelka, A. Agarwal, C. Bogart, Y. Song, M. Sakr, Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses?, arXiv preprint arXiv:2303.09325 (2023).

[14] J. Savelka, A. Agarwal, M. An, C. Bogart, M. Sakr, Thrilled by your progress! large language models (gpt-4) no longer struggle to pass assessments in higher education programming courses, arXiv preprint arXiv:2306.10073 (2023).

[15] S. R. Piccolo, P. Denny, A. Luxton-Reilly, S. Payne, P. G. Ridge, Many bioinformatics programming tasks can be automated with chatgpt, 2023. arXiv:2303.13528.

[16] J. Savelka, A. Agarwal, C. Bogart, M. Sakr, Large language models (gpt) struggle to answer multiple-choice questions about code, 2023. arXiv:2303.08033.

[17] M. Wermelinger, Using github copilot to solve simple programming problems (2023).

[18] T. Phung, J. P. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. K. Singla, G. Soares, Generating high-precision feedback for programming syntax errors using large language

models, ArXiv abs/2302.04662 (2023).

[19] S. MacNeil, A. Tran, D. Mogil, S. Bernstein, E. Ross, Z. Huang, Generating diverse code explanations using the gpt-3 large language model, ICER '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3501709.3544280. doi:10.1145/3501709.3544280.

[20] S. MacNeil, A. Tran, A. Hellas, J. Kim, S. Sarsa, P. Denny, S. Bernstein, J. Leinonen, Experiences from using code explanations generated by large language models in a web software development e-book, SIGCSE 2023, ACM, New York, NY, USA, 2023, p. 931–937. URL: https://doi.org/10.1145/3545945.3569785. doi:10.1145/3545945.3569785.

[21] K. Tan, T. Pang, C. Fan, Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects, 2023. arXiv:2305.03433.

[22] M. Wollowski, Using chatgpt to produce code for a typical college-level assignment, AI Magazine 44 (2023) 129–130. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12086. doi:https://doi.org/10.1002/aaai.12086. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12086.

[23] J. White, S. Hays, Q. Fu, J. Spencer-Smith, D. C. Schmidt, Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design, 2023. arXiv:2303.07839.

[24] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023. arXiv:2303.17580.

[25] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, X. Li, Auggpt: Leveraging chatgpt for text data augmentation, 2023. arXiv:2302.13007.

[26] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[27] Y. Li, M. Rakovic, B. X. Poh, D. Gasevic, G. Chen, Automatic classification of learning objectives based on bloom's taxonomy, in: A. Mitrovic, N. Bosch (Eds.), EDM, International EDM Society, Durham, United Kingdom, 2022, pp. 530–537. doi:10.5281/zenodo.6853191.