Bridging the Measurement Gap: A Large Language Model Method of Assessing Open-Ended Question Complexity

Tuval Raz (tuval.raz@campus.technion.ac.il)
Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology
Haifa 3200003, Israel

Simone Luchini (skl5875@psu.edu)
Department of Psychology, Pennsylvania State University
140 Moore Building, University Park, PA 16802, USA

Roger E. Beaty (rebeaty@psu.edu)
Department of Psychology, Pennsylvania State University
140 Moore Building, University Park, PA 16802, USA

Yoed N. Kenett (yoedk@technion.ac.il)
Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology,
Haifa 3200003, Israel

Abstract

Question-asking, an essential yet understudied activity, holds significant implications for fields such as learning, creativity, and cognitive development. The quality, and complexity in particular, of the questions are recognized as crucial factors affecting these fields. Previous research explored question complexity through Bloom's taxonomy, but measurement remains challenging. Recent advancements have enabled automated scoring of psychological tasks but have not been applied to open-ended question complexity. Here, we address this gap by employing large language model (LLM) techniques to predict human ratings of open-ended question complexity. Our results reveal that our LLM-generated complexity scores correlated strongly with human complexity ratings in both the holdout-responses (r = .73) and holdout-item set (r = .77), whilst also exceeding baseline methods tested. The research emphasizes the significance of LLMs in psychological research and their potential in automating question complexity assessment. This study also highlights exciting possibilities for usage of LLMs in education and psychology.

Keywords: LLM; Bloom Taxonomy; Creativity

Introduction

Question asking is a common and everyday activity which can bridge gaps in knowledge or resolve uncertainty (Ronfard et al., 2018), yet only a very rudimentary understanding of it currently exists (Kearsley, 1976), (but see: Gottlieb, 2021; Sasson & Kenett, 2023). Question asking is central to learning (Chin & Osborne, 2008), whilst also being an important component of educational programs (Chin & Brown, 2002) and creativity (Acar et al., 2023). Nevertheless, few methods exist for automatic scoring of question asking (Jayakodi et al., 2015; Mohammed & Omar, 2020) and to our knowledge, none which have assessed open-ended questions using LLMs, which perform better than previous approaches (Vaswani et al., 2017). Due to the significant role of complexity in question creativity (Raz et al., 2023), in the present study, we extend research on automatic complexity scoring of a creative question asking task, by developing and training a LLM to predict human-rated complexity scores for questions generated in a creative question asking task (Raz et al., 2023).

Open-ended vs. close-ended questions

Ortlieb et al. (2012) argue that the ultimate goal of education should be advancing beyond the use of the closed questioning style towards open-ended questions. Open-ended and close-ended questions differ in several characteristics. Close-ended questions limit the respondent to the answers offered and require engaging in convergent thinking (i.e., converging on a single correct solution), while open-ended questions allow expressing an opinion without being largely influenced by the question designer and engaging in divergent thinking (i.e., diverging on multiple possible solutions). The advantages of open-ended questions include the possibility of discovering spontaneous responses, and avoiding biases that result from suggesting responses, which occur in close-ended questions (Reja et al., 2003).

This is especially pertinent in education as teacher's questions are indispensable components of classroom discourse and student learning (Salmon & Barrera, 2021). Research on teachers' questions reveal that closed-ended questions are used more than open-ended questions in teaching (Çakır & Cengiz. 2016), a practice which has been criticized (Nunan, 1987). Baloche (1994) and Khan and Inamullah (2011) argue that a teacher's ability to ask open-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ended questions is crucial for the development of complex, creative thinking which involves more elaborate and abstract ideas such as the creation of new topics and the expression of opinions. Although complexity seems to be an essential part of question asking, it is not necessarily clear how to best measure and classify it.

Bloom's Taxonomy

One common approach to evaluate question complexity is utilizing the Bloom taxonomy (Bloom et al., 1956), which has been widely accepted as a guideline in designing learning objectives of differing levels of cognitive complexity (Adams, 2015; Goh et al., 2020). Specifically, the taxonomy includes six cognitive levels, which are hierarchically ordered from simple to complex (Krathwohl; 2002). Previous studies have applied the Bloom taxonomy to the evaluation of question complexity: assigning each question a score from one to six and allowing for quantitative analyses to be conducted on question complexity (Plack et al., 2007; Zheng et al., 2008).

Several attempts at using LLMs to predict Bloom taxonomy scores have been made in the past (Gani et al., 2023; Hwang et al., 2023). In one case researchers automated the quality evaluation of multiple-choice questions in introductory chemistry and biology courses (Hwang et al., 2023). This was only partially successful, as model accuracy varied greatly depending on question type. The findings are further complicated as the author's Bloom scores were sourced from a single human rater, raising potential issues of rater subjectivity. In contrast, Gani et al. (2023) developed a Bloom's Taxonomy-based classification approach using an LLM and labeled multiple choice exam questions as training data, achieving good accuracy (86%) compared to previous computational models. The study compared embedding approaches and showed that RoBERTa is the most optimal, and suggested future work could include testing with larger datasets to evaluate its scalability. Critically, multiple choice questions such as the ones used in the study are usually closed-ended, single-solution tasks (SST) (de Vink et al., 2021), which are binary scored for correctness and usually require closed convergent thinking, whereas open-ended questions that require divergent thinking and involve multiple solution tasks (MST) are usually evaluated in terms of fluency, flexibility, and originality. Previous research has indicated that creative thinking is more strongly related to MST than SST performance (de Vink et al., 2021), that asking questions is a key trait of creativity and an integral part of the creative process, and that question complexity is closely related to creativity (Raz et al., 2023). Thus, the need is clear for integrating divergent open-ended question asking together with larger datasets in developing new LLM based approaches to automatically predict question complexity.

Question Asking and Creativity

Asking questions is both a key part of creativity and an important component of the creative process (Acar et al., 2023) that likely facilitates information seeking behavior, and has been shown to be part of the creative problem-solving process in children (Torrance, 1970). Recently, Raz et al. (2023) explored the relation between question asking and creativity using the alternative questions task (AQT). The AQT requires participants to generate creative and unusual questions about common objects (e.g., pen, book, shoe) such as "who invented the first pencil". Responses were rated separately in terms of their creativity, using a 1 (not at all creative) to 5 (very creative) scoring method (Runco & Mraz, 1992; Silvia et al., 2008), and complexity, according to Bloom's taxonomy. They also found that question complexity and creativity were positively related: questions which were higher on the Bloom taxonomy (i.e., more complex) were also scored as more creative, and those that were less complex were scored as less creative. The researchers provided the first proof that question complexity and creativity are related, such that stronger creative abilities will accompany stronger question asking abilities.

Raz et al. (2023) noted that subjective scoring of responses according to the Bloom taxonomy may suffer from the same limitations as subjectively rated creativity scores (Kaufman & Baer, 2012; Silvia et al., 2008), such as inconsistent rater agreement, rater fatigue and high costs. Thus, automating Bloom taxonomy scoring by means of computational approaches may help overcome these limitations. Recent advances in natural language processing (NLP) tools such as semantic distance have allowed for the automated scoring of psychological tasks (e.g., Beaty & Johnson, 2021). This has made it possible to overcome the typical bottlenecks of human scoring (e.g., high labor costs; (Reiter-Palmon et al., 2019). One notable advancement has come in the form of large language models (LLMs). When fine-tuned, these models have demonstrated superior performance on a number of creative thinking tasks (Dumas et al., 2021; Luchini et al., 2023). Thus, several computational approaches exist for automating Bloom taxonomy scoring such as using NLP semantic distance tools or newer Language model methods (Stevenson et al., 2022).

Large Language Models in psychological research

LLMs are computational tools that are used for a variety of tasks involving language data (Vaswani et al., 2017). They are a class of deep neural networks that undergo pre-training on large amounts of text data, for the purpose of learning and generating language. These NLP models have found success in psychology given the widespread importance of language data in research (Demszky et al., 2023). For example, LLMs can been used to generate experimental stimuli (Laverghetta & Licato, 2023), model word learning across the lifespan (Portelance et al., 2020), predict personality traits from text (Peters & Matz, 2023), and predict responses to problem solving tasks (Luchini et al., 2023).

LLMs are typically pre-trained through unsupervised learning, a training procedure which involves automatic pattern detection from unlabeled data—text that is not assigned any tag or number that the model has to predict. For LLMs this takes the form of iterative word prediction problems, where the model is required to predict a missing word from the context or vice versa (Jiang et al., 2020). In this way, LLMs can manipulate language efficiently and are thus able to outperform previous NLP tools on tasks they were never trained on (Vaswani et al., 2017).

The present research

The present study aimed to address the gap in the literature on questions asking and the role of complexity by developing an LLM model capable of scoring participant's open-ended questions according to the Bloom taxonomy. This was done in an attempt to advance the availability, cost effectiveness and reliability of question complexity and creativity scoring and to highlight the advantages of the usage of LLMs in education and psychology and their potential in helping study how we ask creative questions.

The model was trained on data we compiled from thousands of human-rated responses to the alternative questions task (AQT), a creative question asking task introduced by Raz et al. (2023). Responses were questions asked about everyday objects spanning a total of 6 items taken from the suggested items provided by Beaty et al. (2022). To evaluate model performance, its predictions were compared to three other scoring methods—elaboration (i.e. word count) and two semantic distance methods (MAD & DSI) —which reliably predict human creativity ratings in divergent thinking tasks (Luchini et al., 2023).

Materials and Methods

Participants

Data analyzed comes from two different sources. The first is reanalysis of data collected by Raz et al. (2023) (47.9% male, 50.4% female, 1.7% preferred not to say; mean age = 26.1 years, SD = 6.41 years). The second dataset is recent data collected for a larger ongoing study on how priming question asking capacity impacts creative problem solving (49% male, 50% female, 1% preferred not to say, mean age = 29.35 years, SD = 9.62 years). Both samples correspond to each other and are composed of data collected from participants recruited on Prolific Academic (N = 723). The total dataset consisted of 10,282 responses to the alternative questions task (AQT), spanning a total of 6 items (pencil, sock, pillow, purse, clock, and knife). Average number of AQT responses per participant was (M = 4.74, SD = 2.341).

Methods

Alternative Questions Task (AQT)

The AQT requires participants to generate creative and unusual questions about three common objects in two minutes for each object (Raz et al., 2023). The AQT was developed to test open-ended divergent question asking in general, beyond the classroom context (Raz et al., 2023). AQT objects were taken from the suggested items provided by Beaty et al. (2022) for the alternative uses task (AUT) which the AQT is based on (pencil, sock, pillow, purse, clock, and knife). Participants were explicitly instructed to come up with as many original and creative questions for objects as they can (Said-Metwaly et al., 2020). Creative questions were defined in the study as questions that strike people as clever, unusual, innovative, or different. Participants provided their responses on a single page with 30 available input fields. Time limits were two minutes per object.

Bloom Taxonomy. AQT responses were externally scored for their respective Bloom level (from one to six: *Remember*, *Understanding*, *Applying*, *Analyzing*, *Evaluating* and *Creating*). The revised edition of Bloom's taxonomy (Krathwohl, 2002) was used. Online raters from Prolific Academic were instructed on the Bloom taxonomy and subjectively rated AQT responses by assigning the level they ascertained from the response. Rating instructions included an explanation of the types of questions asked for each Bloom level, alongside key terms related to each level and examples of scoring. Each object cue was rated by ten different independent raters for study 1 (Raz et al., 2023) due to some raters failing attention checks, and by five raters for the priming questions dataset. Raters who failed attention checks or gave incomplete ratings were excluded from the final dataset. Reliability metrics for AQT objects on their Bloom level ratings were overall good (Ko & Lee, 2016) and as follows: pencil (N = 4, α = .752), pillow (N = 3, α = .720), sock (N = 10, α = .768) , knife (N =5, α = .702), purse (N = 3, α = .63), and clock (N = 4, α = .61).

Automated Originality Scoring

Given the extensive work done on validating semantic distance as a measure of originality (e.g., Patterson et al., 2023), both MAD and DSI scores can serve as a good tool or baseline against which to compare our LLM.

Maximal Associative Distance (MAD). We computed MAD scores (Yu et al., 2023) for all AQT responses following the approach described by Patterson et al. (2023). MAD was selected given it is an unsupervised machine learning approach (i.e., human-rated originality scores are never shown to the model) for extracting originality measures from text data. The MAD method has been shown to outperform previous compositional techniques at predicting human-rated originality and to correlate with Bloom scores on the AQT (Raz et al., 2023). Semantic distance scores are first computed between all words in a response (e.g., *can I bend the pencil without breaking it*) and the prompt (e.g., *pencil*). Then, only the most distant word is retained. Response-level MAD scores are thus the semantic distance of the most distant word from the prompt.

Divergent Semantic Integration (DSI). We calculated DSI scores for AQT responses as a baseline against which LLM performance was compared to as this is also an unsupervised machine learning technique for extracting originality measures from text data. As such, it is a good benchmark to evaluate the performance of supervised models which involve the fine-tuning of model weights. DSI significantly predicts originality on other creativity tasks involving long-form text responses (DiStefano et al., 2024; Johnson et al., 2023, Luchini et al., 2023). To calculate DSI scores, word embeddings from the BERT model were first extracted from a pre-trained machine learning model. These embeddings were then used to calculate the semantic distance between all pairs of words in a response.

Bloom scoring LLM

Model. The RoBERTa base model was selected for fine-tuning. RoBERTa constitutes an improvement on the BERT model (Liu et al., 2019), and was released by Google in 2018 (Devlin et al., 2018). RoBERTa is a transformer model (see Vaswani et al., 2017) which underwent self-supervised training, without any human labeling being presented. The architecture of the model is similar to BERT with some changes in the pretraining procedure, including an increased amount of data it was trained on (Liu et al., 2019). Given its smaller size, this base version comes with a reduced computational cost compared to larger versions of the same model. This version has 123 million parameters (i.e., weights). RoBERTa is a LLM that was pre trained with a bidirectional approach (i.e., the model saw entire sentences when making predictions) applied to context leveraging (i.e., filling in the blanks by drawing from the surrounding context). Of note, RoBERTa has been shown to perform strongly on a variety of linguistic tasks (Gillioz et al., 2020). The model is available on the open-access Huggingface platform (https://huggingface.co/). For fine-tuning, the "Huggingface Transformers" suite of the *PyTorch* package was used via the Python programming language.

Datasets. Data (AQT responses collected as described above) was randomly split into training, validation, and holdout (response and items) sets following a 70/10/20 ratio. The *training data* was employed to fine-tune the model, as the model saw both responses and human Bloom ratings and learned to predict the ratings. The *validation data* served the purpose of iteratively testing different variations of the model, each trained with different combinations of hyperparameter values, to determine the best settings. The *holdout-responses* data contained responses that the model was never presented during training and allowed the testing of model performance on unseen responses. Responses in the holdout-responses set were associated with AQT items that the model saw during training, and as such served as a test of near-transfer. The *holdout-item* data also contained responses that the model wasn't shown during training, except they were related to an AQT item that the model never saw during training. As such, the holdout-item set was employed as a test of far-transfer of model performance.

The training data was used to adjust the weights of the model. To achieve this, batches of responses from the training set were inputted to the model during training. The model would then predict a single creativity value associated with each response, and the mean squared error between these predicted values and the true human-rated Bloom scores determined the degree of weight adjustment. The validation set was also employed during training but was instead used for the hyperparameter search. It therefore did not serve the purpose of adjusting model weights. Model predictions for the validation set were compared across a variety of model settings, retaining only the best combination for later testing. The responses in the holdout-responses set were only presented to the model once the training procedure had finished. This allowed for a test of model generalizability, evaluating whether model predictions could extend to unseen data. We further withheld responses to the item 'clock' from the splitting procedure and assigned them to a holdout-item set. This item was selected for the holdout-item set as it was associated with the smallest number of responses in the entire dataset, compared to the other items, leaving more data to be assigned to the other sets. By evaluating model predictions for the holdout-item set it is then possible to determine generalizability to unseen items. Ideal model performance would involve strong predictions for both unseen responses and items, as this would indicate that the model can be extended to different responses and items.

Hyperparameter Search. Hyperparameters are settings that determine the learning of a model and which are set prior to training. We implemented a hyperparameter search over the number of epochs, the learning rate, the training batch size, and the evaluation batch size based on similar applications (e.g., DiStefano et al., 2024; Luchini et al., 2023).

The number of epochs determines how many times the model will iterate over the entire training dataset during training. For this, we searched between a range of 100 and 130 epochs. The learning rate determines the speed at which the model will learn from the data. It effectively modifies the impact that one batch of data will have on the weights of the model. We searched between learning rate values of 5e-05 and 5e-04. The batch size refers to the number of responses that are inputted to the model during each iteration. Batch sizes were searched separately for the training and the evaluation data (i.e., validation, holdout-responses, and holdout-prompt sets). For both training and evaluation batch sizes, we searched through three possible values of 8, 16, and 32. Low batch values were selected because larger values lead to poor generalizability of model predictions (Keskar et al., 2016). Total model training time took approximately 72 hours.

Hyperparameter search was run by employing the Optuna package in Python language (Akiba et al., 2019). Optuna allows for the search over a variety of hyperparameter combinations by means of a Tree-Structured Parzen Estimator, a Bayesian optimization method. The best hyperparameter settings were identified and used in the training of our final model. All other hyperparameters were left to the default settings for RoBERTa-base. Data, analysis scripts and weights for our final model are available online: (https://osf.io/823ak/?view_only=2f8f7a94ca5e4d57b810afc26054a6d3).

Results

Descriptive analysis

We computed a series of descriptive statistics for the AQT. Across all items, the mean word count was 6.58, and the mean Bloom rating was 2.85. We calculated the intra-class correlation (ICC; Shrout & Fleiss, 1979) between raters in the study and found a strong reliability across all items in the dataset (ICC = 0.76 [95% CI: 0.75, 0.77]) on the average ratings using a two-way random-effects model with absolute agreement. We then computed a series of Pearson's correlations between our variables of interests, separately for each set. Before computing any of the linear regression models, we removed outliers from all variables by excluding datapoints that lay more than 3 standard deviations above or below the mean.

We then calculated correlations between DSI and MAD and found moderate correlations for the training set, r = .54, validation set, r = .47, holdout-responses-set, r = .53 and holdout-item set, r = .61 (all p's < .001). Next, we then calculated correlations between MAD and word count. Strong to moderate correlations were observed throughout all sets, with r = .58 for the training set, r = .51 for the validation set, r = .57 for the holdout-responses set, and r = .68 for the holdout-item set (all p's < .001).

We then calculated correlations between DSI and word count. Strong correlations were observed throughout all sets, with r = .70 for the training set, r = .66 for the validation set, r = .69 for the holdout-responses set, and r = .73 for the holdout-item set (all p's < .001).

Finally, we computed correlations between baseline measures and human-rated Bloom scores. For word count, we observed moderate correlations of r = .40 for the training set, r = .30 for the validation set, r = .44 for the holdout-responses set, and r = .37 for the holdout-item item set. For MAD, we observed moderate correlations of r = .31 for the training set, r = .22 for the validation set, r = .32 for the holdout-responses set, and r = .42 for the holdout-item set. Finally, for DSI, we observed moderate correlations of r = .39 for the training set, r = .27 for the validation set, r = .39 for the holdout-responses set, and r = .34 for the holdout-item set (all p's < .001).

LLM Prediction of Bloom Ratings

Hyperparameter settings for our final RoBERTa model included 114 epochs, a learning rate of 9.2e-05, and a batch size of 16 for both training and evaluation. The fine-tuned RoBERTa model was then used to predict Bloom scores for each response. These predicted scores were then included in a series of linear regressions against the human rated scores (**Figure 1**). We found that our model perfectly predicted the human ratings at r = .99 on the training set, and strongly predicted them on the validation set r = .76 (all p's < .001).

As a test of model generalizability, we further explored correlations for the holdout-responses set. The held-out test set consisted of responses that were neither used for model selection nor seen by the model during its training. It thus served as a test of the model's ability to generalize to responses it wasn't trained on. We found that the correlation between model predictions and the holdout-responses was substantially larger, r = .73, p < .001, than baseline measures. The results demonstrate that fine-tuned LLMs, can strongly capture and predict human creativity ratings of question complexity. We then tested the correlation with the holdout-item set, which consisted of item inputs that were neither used for model selection nor seen by the model during training. We found a substantially larger, r = .77, p < .001, correlation than baseline measures. This highlights how LLMs can accurately predict human ratings of question asking tasks and offer a reliable and efficient alternative to labor-intensive and subjective human ratings of question complexity.

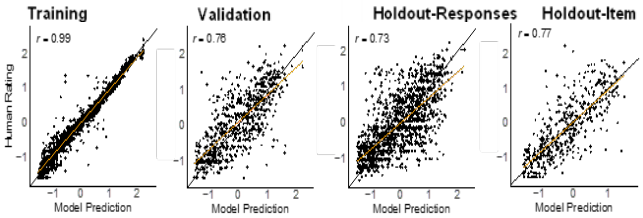

Figure 1: Linear regressions between human-rated Bloom scores and model predictions. Single responses are denoted by the black dots. Ideal performance (r = 1) is denoted by the black line, while the orange line is the line of best fit. All p's < .001.

Discussion

Questions play a critical role in learning, education, and creativity. However, much is still unknown about the role of asking complex question in cognition. This is in part due to the challenge of scoring and assessing open-ended questions, an issue that is also relevant in broader creativity research (Kaufman, 2019). Despite recent advancements in LLMs and their emerging role in psychological research (Zhang et al., 2023), little research has examined automated scoring of question asking (Gani et al., 2023; Hwang et al., 2023). Furthermore, to our knowledge, none has been applied to scoring open-ended questions.

The current study capitalized on advanced LLMs to develop automatic and accurate scoring of open-ended question complexity, based on the Bloom's taxonomy (Bloom et al., 1956). We trained a LLM (RoBERTa) to predict human-rated Bloom scores for responses to the AQT, which measures creative question asking (Raz et al., 2023). Bloom scores are a measure of the cognitive complexity of an AQT response, subjectively scored by human raters (with high inter-rater agreement).

Our fine-tuned LLM demonstrated robust predictions of Bloom scores, surpassing those achieved by semantic distance models or word count. The model generalized its predictions beyond the data it was trained on and demonstrated good performance on the test data, reaching a Pearson correlation above .75. Of note, strong predictions were also observed for an unseen prompt item. By evaluating model predictions for the unseen prompt item, it is then possible to determine generalizability of the model. Ideal model performance would involve strong predictions for both unseen responses and items. We provide the first evidence that a LLM can robustly predict Bloom complexity scores and automatically score open-ended questions.

Model predictions strongly correlated with human-rated Bloom scores in both the holdout-responses and holdout-item set. Model performance on the holdout-item set was even slightly better than inter-rater agreement between human raters in the study. Correlations between the baseline measures and human-rated Bloom scores were consistently moderate for both holdout responses and item sets. Thus, the model substantially outperformed baseline measures. These findings indicate that the LLM model was able to pick up a big part of how humans evaluate questions and was able to re-apply this knowledge to new data.

Question asking is an important human capacity, related to curiosity, problem finding and information seeking behavior (Kenett et al., 2023; Raz & Kenett, 2024; Raz et al., 2023). But critically, not all questions are the same. The type of questions used can have a very important role in constructing a facilitative environment for information seeking, education or higher-level thinking (Çakır & Cengiz, 2016).

Çakır and Cengiz (2016) support the idea that open-ended questions elicit more utterances from students, enhance creativity, and encourage the learner to contemplate and explore. Conversely, close-ended questions limit the respondent to the set of alternative answers offered in the question and bias thinking towards them. The model developed in this study aims to further advance research into this area of open-ended questions by adding an additional tool to the limited but expanding toolset of question asking measurement.

There are some potential limitations concerning the results of this study. As noted by Luchini et al. (2023), smaller, older LLMs such as RoBERTa and GPT-2 have been shown in the past to underperform on certain benchmark tasks when compared to newer, more advanced ones. The current analysis should therefore be extended to larger models, such as GPT-4, to evaluate whether predictions can be further improved. However, larger models like GPT-4 are currently not freely available, and researchers would incur costs when employing these models. Additionally, the model developed in this study was trained on averaged continuous scores of Bloom complexity via a regression model. As such, the output scores of the model are continuous, but are rounded to the nearest whole level in order to display familiar bloom levels. This is in contrast to a classification model which outputs discrete scores, in this case corresponding to the one to six Bloom levels. We opted to use a prediction, regression-based approach to align with previous LLM applications of open-ended responses (e.g., Distefano et al., 2024; Luchini et al., 2023). The question of whether the complexity of questions is discrete or continuous is still an open one and requires further research into how we measure complexity. Despite these limitations, the results suggest several theoretical and practical implications.

Research with elementary and college aged students has shown that they can quickly be taught how to ask higher level questions rather than lower complexity factual questions, and that this leads to improvements in learning and reading comprehension (Ronfard et al., 2018). It is therefore possible that educators may greatly benefit from automated methods of assessing open-ended question complexity, thus helping to foster creativity, learning and advanced comprehension in students. Future research should also focus on expanding the arsenal of automated psychological tests using LLMs, which will greatly improve the accessibility of these tests, and on creating online "one stop shop" resources combining many automated scoring techniques similarly to the work done by Beaty and Johnson (2021).

Conclusion

In this study, we introduce a novel approach to automatically score the Bloom taxonomy complexity of open-ended questions using a fine-tuned LLM. Our results reveal that LLM-generated Bloom scores correlated strongly with human ratings—greatly exceeding baseline measures. These results highlight the unique ability of LLMs to accurately predict ratings of open-ended questions. Our study offers a reliable and efficient alternative to labor-intensive and subjective human ratings of question complexity—improving the reproducibility and scalability of complexity assessment. This study also emphasizes the exciting potential for the continued usage of LLMs in education and psychology and the possibilities they unlock in studying how we ask creative questions about the world and help us build the educators and pedagogical programs of the future.

Acknowledgments

R.E.B. is supported by grants from the National Science Foundation [DRL-1920653; DUE-2155070]. This work was partially supported by the US-Israel Binational Science Fund (BSF) grant (number 2021040) to R.E.B and Y.N.K.

References

Acar, S., Berthiaume, K., & Johnson, R. (2023). What kind of questions do creative people ask? *Journal of Creativity,* 100062. https://doi.org/10.1016/j.yjoc.2023.100062

Adams N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association*, *103*(3),152–153. https://doi.org/10.3163/1536-5050.103.3.010

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Albergaria-Almeida, P. (2011). Critical thinking, questioning and creativity as components of intelligence. Procedia - *Social and Behavioral Sciences, 30*, 357–362. https://doi.org/10.1016/J.SBSPRO.2011.10.070

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Baloche, L. (1994). Breaking down the walls. *The Social Studies.85*.25-30. https://doi.org/10.1080/00377996.1994.10118776

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02529

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, *34*(3), 245-260. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences,* 27(7), 671-683. https://doi.org/10.1016/j.tics.2023.04.004

Bloom B. S., Krathwohl D. R., & Masia B. B. (1956). *Taxonomy of educational objectives: the classification of educational goals*. David McKay Company.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

Çakır, H. and Cengiz, Ö. (2016) The use of open ended versus closed ended questions in Turkish classrooms. *Open Journal of Modern Linguistics*, *6*, 60-70. doi: 10.4236/ojml.2016.62006.

Chin, C., & Brown, D. E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education, 24*(5),521–549.https://doi.org/10.1080/09500690110095249

Chin, C., & Osborne, J. (2008). Students' questions: a potential resource for teaching and learning science. *Studies in Science Education, 44*(1), 1–39. https://doi.org/10.1080/03057260701828101

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113-118.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language *ArXiv*. https://arxiv.org/abs/1810.04805v2

DiStefano, P. V., Patterson, J. D., & Beaty, R. (2024). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal,* 1-15. https://doi.org/10.1080/10400419.2024.2326343

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15(4), 645–663. https://doi.org/10.1037/aca0000319

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930–1955. *Studies in Linguistic Analysis*. Oxford, UK: Blackwell.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, 23, 129-139. https://doi.org/10.101t6/j.tsc.2016.12.005

Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies, 28*(12), 15893–15914. https://doi.org/10.1007/s10639-023-11842-1

Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the Transformer-based models for NLP tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179-183). IEEE.

Goh, T.T., Mohamed, H., Jamaludin, N.A., Ismail, M.N., & Chua, H.S. (2020). Questions classification according to Bloom's taxonomy using universal dependency and Word Net. *Test Engineering and Management. 82*. 4374-4385

Gottlieb, J. (2021). The effort of asking good questions. *Nature Human Behaviour, 5*(7), 823-824. https://doi.org/10.1038/s41562-021-01132-6

Grévisse, C. (2024). Comparative quality analysis of GPT-based multiple choice question generation. In H. Florez & M. Leon (Eds.), *Applied Informatics* (pp. 435–447). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-46813-1_29

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hardy, M., Sucholutsky, I., Thompson, B., & Griffiths, T. (2023). Large language models meet cognitive science: LLMs as tools, models, and participants. *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*(45). https://escholarship.org/uc/item/6dp9k2gz

Hwang, K., Challagundla, S., Alomair, M., Chen, L. K., & Choa, F. S. (2023). Towards AI-assisted multiple choice question generation and quality evaluation at scale: Aligning with Bloom's taxonomy. *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*, 1--8.

Jayakodi, K., Bandara, M., & Perera, I. (2015). An automatic classifier for exam questions in engineering: A process for Bloom's taxonomy. *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 195-202.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, *8*, 423-438.

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, *55*(7), 3726–3759. https://doi.org/10.3758/s13428-022-01986-2

Kaufman, J. C. (2019). Self-assessments of creativity: Not ideal, but better than you think. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 187-192. https://doi.org/10.1037/aca0000217

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal, 24*(1), 8391. https://doi.org/10.1080/10400419.2012.649237

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S.(2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts, 7*(4), 332–340. https://doi.org/10.1037/a0034809

Kearsley, G. P. (1976). Questions and question asking in verbal discourse: A cross-disciplinary review. *Journal of Psycholinguistic Research, 5(*4), 355–375. https://doi.org/10.1007/BF01079934

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*,*27*,11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *ArXiv*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *ArXiv*. https://doi.org/10.48550/arXiv.2205.11916

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*, 212-218. http://dx.doi.org/10.1207/s15430421tip4104_2

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Laverghetta, A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (EEA 2023),* 414-428. https://doi.org/10.18653/v1/2023.bea-1.34

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151-171.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewvis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *Arxiv*. https://doi.org/10.48550/arxiv.1907.11692

Luchini, S., Maliakkal, N. T., DiStefano, P. V., Patterson, J. D., Beaty, R., & Reiter-Palmon, R. (2023). Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings. *PsyArxiv*.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv.*

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS ONE*, *15*(3), e0230442. https://doi.org/10.1371/journal.pone.0230442

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979

Nunan, D. (1987). Communicative language teaching: making it work. *ELT Journal, 41*, 136-145. http://dx.doi.org/10.1093/elt/41.2.136

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 101356.

Ortlieb, E., Bowden, R., Inman, A., Hu, B. Y., Pate, R. S., Gauthier, L. R., & Schorzman, E. M. (2012). *Educational Research and Innovations*. CEDER, Texas A&M University-Corpus-Christi. https://hdl.handle.net/1969.6/97734

Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2023). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02258-3

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S. Baas, M., Baker, B. S., Barbot, B., Benedek, M., Chen, Q., Corazza, G. E., Forthmann, B., Karwowski, M., Kreisberg-Nitzav, A. Kenett, Y. N., Lubart, T., Mercier, M., Miroshnik, K., Ovando-Tellez, M., Primi, R., Puente-Diaz, R., Said-Metwaly, S., Stevenson, C., Volle, E., van Hell, J. G., & Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, 17(4), 495-507. DOI: https://doi.org/10.1037/aca0000618

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Peters, H., & Matz, S. (2023). Large language models can infer psychological dispositions of social media users. *ArXiv*.

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing reflective writing on a pediatric clerkship by using a modified Bloom's Taxonomy. *Ambulatory Pediatrics: The Official Journal of the Ambulatory Pediatric Association, 7*(4), 285–291. https://doi.org/10.1016/j.ambp.2007.04.006

Portelance, E., Degen, J., & Frank, M.C. (2020). Predicting age of acquisition in early word learning using recurrent neural networks. *Annual Meeting of the Cognitive Science Society*.

Rathje, S., Mirea, D. -M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *PsyArxiv*. https://doi.org/10.31234/osf.io/sekf5

Raz, T., & Kenett, Y. N. (2024). Question asking as a mechanism that facilitates seeking of information [Peer commentary on Ivancovsky, T., Baror, S., & Bar, M. (2023). A shared novelty-seeking basis for creativity and curiosity]. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X23002807

Raz, T., Reiter-Palmon, R., & Kenett, Y. N. (2023). The Role of asking more complex questions in creative thinking. *Psychology of Aesthetics, Creativity, and the Arts.* https://doi.org/10.1037/aca0000658

Reilly, J., Finley, A. M., Litovsky, C., & Kenett, Y. N. (2023). Bigram semantic distance as a measure of conceptual transitions in continuous natural language: Theory, tools, applications. *Journal of Experimental Psychology: General*, 152(9), 2578-2590. **https://doi.org**10.1037/xge0001389

Reimers, N., & Gurevych, I. (2019). Sentence-Bert: Sentence embeddings using siamese Bert-networks. *ArXiv*.

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in Web questionnaires. *Developments in Applied Statistics*, *19*, 159–177.

Ronfard, S., Zambrana, I. M., Hermansen, T. K., & Kelemen, D. (2018). Question-asking in childhood: A review of the literature and a framework for understanding its development. *Developmental Review, 49*, 101–120. https://doi.org/10.1016/j.dr.2018.05.002

Runco, M. A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement, 52*(1), 213–221. https://doi.org/10.1177/001316449205200126

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., & Van den Noortgate, W. (2020). Testing conditions and creative performance: Meta-analyses of the impact of time limits and instructions. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1), 15-38. https://doi.org/10.1037/aca0000244

Salmon, A. K., & Barrera, M. X. (2021). Intentional questioning to promote thinking and learning. *Thinking Skills and Creativity*, *40*, 100822. https://doi.org/10.1016/j.tsc.2021.100822

Sasson, G., & Kenett, Y. N. (2023). A mirror to human question asking: Analyzing the Akinator online question game. *Big Data and Cognitive Computing, 7*, 26. https://doi.org/10.3390/bdcc7010026

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Silvia, P. J. (2008). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). *Creativity Research Journal*, *20(1)*, 34–39 https://dx.doi.org/10.1080/10400410701841807

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). Putting GPT-3's creativity to the (alternative uses) test. *ArXiv*.

Torrance, E. P. (1970). Group size and question performance of preprimary children. *The Journal of Psychology: Interdisciplinary and Applied, 74*(1), 71–75. https://doi.org/10.1080/00223980.1970.10545279

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv*. http://arxiv.org/abs/1706.03762

Wei, J. M., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Arxiv*.

Yu, Y., Beaty, R. E., Forthmann, B., Beeman, M., Cruz, J. H., & Johnson, D. (2023). A MAD method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts*.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *ArXiv.*

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Assessment. Application of Bloom's taxonomy debunks the "MCAT myth". *Science*, 319(5862), 414–415. https://doi.org/10.1126/science.1147852