Automatic Question Generation and Constructed Response Scoring in Intelligent Texts

Wesley Morris[1,†], Joon Suh Choi[1], Langdon Holmes[1], Vaibhav Gupta[1] and Scott Crossley[1]

[1] *Vanderbilt University, 2201 West End Ave, Nashville, TN 37235*

Abstract

Student-generated constructed responses (i.e. through open response problems) are known to have a significant impact on reading and learning outcomes, but these types of tasks can be difficult to create and time-consuming to score. In this study we experiment with a number of generative and encoder-only large language models (LLMs) to create a Natural Language Processing (NLP) pipeline for automatically generating questions and scoring short constructed responses to those questions. Our pipeline is created as a component of a larger framework for intelligent texts. We present the steps taken in testing the LLMs to develop the pipeline, as well as the results from a preliminary study using the constructed response pipeline on human participants interacting with the Intelligent Texts for Enhanced Lifelong Learning (iTELL). We find that GPT-3.5 is effective at generating questions and reference correct answers. For scoring the constructed responses, we use two encoder models: BLEURT and MPNet. In our trial study, participants report positive experiences with the constructed response task, while giving suggestions about the accuracy and clarity of the feedback.

Keywords

Automatic Constructed Response Scoring, Large Language Models, Automatic Question Generation, Natural Language Processing

1. Introduction

Constructed response items, in which students are prompted to provide an open-ended response to a question [31], are commonly used to assess reading comprehension because of their capacity to encourage active processing of information [33] and to improve learning relative to simply reading [32]. However, scoring of constructed responses can take considerable time and resources [14]. Recent advances in large language models (LLMs) have enabled the possibility of generating and scoring constructed response items automatically and at scale [29, 48]. Automatic generation of questions and automatic scoring of constructed responses to those questions using LLMs could have a significant impact on the capacity to deploy these types of constructed response questions.

This study is part of a broader project to develop the Intelligent Texts for Enhanced Lifelong Learning (iTELL) framework. Intelligent texts, such as those generated by iTELL, are unlike static texts in that they are interactive and dynamically personalized to their users [6]. The iTELL framework ingests static texts on any content domain using a custom content preparation system that guides content editors through the process of formatting the text and reviewing AI-augmented content prior to publication. Text content in iTELL is augmented with keyphrase generation (used for writing feedback) and short answer question generation. The resulting webapp includes a number of interactive features, including annotation and highlighting, summary writing with automatic scoring [36], in-browser Python coding exercises, and the short constructed response items described in the current work.

The purpose of iTELL is to allow teachers, administrators, supervisors, and other content creators to automatically generate intelligent texts on any subject area. The context of this study and its incorporation into the larger iTELL framework require a question generation pipeline based on LLMs with the opportunity for human intervention (i.e., human-in-the-loop [50]), as well as a response scoring pipeline based on LLMs that is fully automated. The pipeline must be content agnostic as well.

The twin necessities for the pipeline to be both automated and content agnostic creates a unique machine learning and natural language processing (NLP) challenge that requires LLMs. Our pipeline can be divided into three parts: automatic question and reference answer generation, review and revision by a content editor, and automatic scoring of the constructed responses generated by users. We performed experiments with several LLMs and NLP methodologies for each task, which are described below. We also tested the pipeline with a group of participants in an online computer science course.

2. Background

The act of constructing knowledge rather than simply receiving it has been researched by academics interested in memory and education for many years. Early work by Jacoby [22] demonstrated that participants who were prompted to construct solutions to an orthographic task were twice as likely to recall the solutions than participants who simply read them. Working contemporaneously, Slamecka and Graf [42] found that memory for words generated by participants was greater than those simply read across several variations of the word recall task. Slameck and Graf refer to the phenomenon wherein

Leveraging Large Language Models for Next Generation Educational Technologies, July 14, 2024, Atlanta, Georgia, USA

wesley.g.morris@vanderbilt.edu (W. Morris)

memory is improved through construction as the 'generation effect'. Subsequent research has confirmed that the generation effect is a stable and significant construct in learning [2, 7, 10, 34]. One meta-analysis of 86 studies indicated that construction may improve learning by almost half a standard deviation [4], while a later meta-analysis of 310 experiments suggested that construction led to a 10 percentage-point increase in learning over reading [32].

Given the robust and well-documented impact of the generation effect on learning, especially when it is compared to just reading, it makes sense to examine how the generation effect can be leveraged to improve reading comprehension. Early research on reading comprehension questions focused on classifying questions according to Bloom's Taxonomy [1], with lower-order skills including recall and paraphrasing, while higher-order inference skills include analysis, application, synthesis, evaluation, and creation [39]. Research on in-class reading comprehension questions found that the type of questions asked by teachers had a strong effect on the linguistic complexity of the students' responses, with high-challenge (*wh*-word questions) eliciting more complex responses [5].

Research has also indicated that constructed response items elicit higher-level processing than multiple choice items [23], and there is robust evidence that tasks with lower constraints, such as constructed response tasks, more effectively elicit the generation effect than tasks with higher constraints, such as multiple-choice questions [16, 32].

Automatic short constructed-response question generation has a long history in the literature [28] and recent research has shown that generative models such as GPT can produce most question types with statistically significant validity [29, 48]. Until recently, however, constructed response items have been considered too time-consuming and expensive to score [14], making them difficult to implement in the classroom. In fact, given the cost and complexity of grading constructed response items, some researchers have recommended phasing them out altogether [20]. However, recent advancements in artificial intelligence, specifically LLMs, have made it possible to score these types of items automatically at very low cost in time and money.

Early work in automated scoring of constructed response items required content creators to employ content experts to manually define lists of target concepts [44, 45]. Alternatively, techniques such as latent semantic analysis have been used, provided that the model was provided with correct answers or hints [9]. With the advent of transformer-based LLMs [47] and especially encoder-only models such as Bidirectional Encoder Representations from Transformers (BERT) [11], automated scoring of constructed responses became attainable. For example, Jung, Tyack, and Davier [24] trained a neural network to predict human scores of four items, while researchers in another study [37] were able to train DeBERTa [19] to score ten math items. However, in both studies large training sets were required and the models were unable to generalize to items other than the items they were trained on.

While previous research has used deep learning methods to attain high scoring accuracy when given very large, labelled training sets of the items to be scored, our goal is to create a general, domain agnostic pipeline for automatic question generation and automatic constructed response scoring for use in iTELL. Our constructed response pipeline includes both question generation and the scoring of constructed responses to those questions. As such, this study is guided by three research questions:

RQ1: Can LLMs be leveraged to automatically generate constructed response questions and correct responses within iTELL?

RQ2: Can open source, encoder-only LLMs be utilized to automatically score constructed responses in iTELL?

RQ3: What are users' experiences with automated constructed response generation and scoring feedback informed through LLM in the context of iTELL?

3. Study 1: Constructed Response Generation

3.1. Method

The first phase of the constructed response pipeline is question generation, which takes place in iTELL's content preparation system. During content development, a content editor segments each page of a text into "chunks" of text, or several paragraphs on a specific subtopic, which are stored in a database. Some of these chunks, such as learning objectives, glossaries, and exercises are disregarded for the purposes of constructed response generation. For the remaining chunks, iTELL generates constructed response questions during content creation. We experimented with GPT-3.5 for generating constructed response questions and answers, which was state-of-the-art at the time of design.

Unlike the original transformer architecture described by Vaswani et al. [47], which included both encoder and decoder blocks, the Generative Pretrained Transformer (GPT) family of models are auto-regressive, meaning that they predict the next sequence of tokens using only the tokens already provided [17, 46]. The power of GPT comes from its massive parameterization (117 million parameters for GPT-1 up to 175 billion for GPT-3 [25]) and its proprietary training set. The training process includes Reinforcement Learning through Human Feedback (RLHF), wherein a reward model is trained on human preferences, then the reward model is used as a reward function to finetune the model [38]. GPT-3.5 achieves good results in benchmark tests [27], suggesting that it might be appropriate for our purposes.

To determine the accuracy of LLMs in generating constructed response questions based on a source text, we used GPT-3.5. Our source text was the Principles of Macroeconomics 2nd Edition textbook available from OpenStax [18]. This electronic text comprises 523 subsections with a mean length of 332.96 words (SD = 236.38). We first randomly sampled 20 subsections and

generated questions of three different question types (recall, summary, and inference) for each subsection using GPT-3.5. In total, we generated sixty questions using a simple prompting strategy of providing the passage and prompting the model to generate a question based on the passage of the appropriate question type. The questions were scored by human raters on a three-point scale {0, 1, 2}. A score of 0 indicates that the question was either incoherent or unrelated to the context. A score of 1 indicates that the question was coherent and related to the context, but of the wrong question type. A score of 2 indicates that the question was coherent, related to the context, and of the correct question type.

We next tested GPT's ability to generate correct reference answers. The answers were also rated by human raters using a binary scale, {0, 1}, with score of 0 indicating that the answer was incoherent or wrong and a score of 1 indicating that the answer was correct. In both cases, each question or answer was scored by two raters with a third rater arbitrating in cases of disagreement.

3.2. Results

The human scores for the constructed response questions developed by GPT-3.5 reported a Cohen's Kappa of 0.65, representing reasonable interrater reliability. Ratings indicated that GPT-3.5 generated questions that were coherent and related to the source in 100% of cases. The questions had an average length of 12.48 words (sd = 4.63). However, 35% (n = 21) of the 60 generated questions were found to be of the wrong question type, with the model primarily generating recall questions regardless of how it was prompted.

For the reference answers to the questions, raters reported perfect agreement for GPT-3.5, with both raters finding that all answers were coherent and correct. The answers had a mean length of 16.13 words (sd = 9.90). As a result of this study, we decided to use GPT-3.5 for the generation of both constructed response questions and correct reference answers within iTELL for the constructed response tool. However, because of the lower accuracy in generating questions of the correct question type, we decided to prompt GPT-3.5 to generate questions without specifying the question type.

4. Study 2: Automatic Scoring of Constructed Responses

4.1. Method

4.1.1. Dataset

Our goal is to have an automatic scoring model that can discriminate between correct and incorrect constructed responses based on a source context. To accomplish this, we trained automatic scoring models using the dataset of questions and answers found in the Multi-Sentence Reading Comprehension (MultiRC) dataset. MultiRC [26] is a dataset of multiple-choice reading comprehension questions, answers, and sources. The sources include news articles, Wikipedia articles, elementary school science textbooks, and fiction. We chose MultiRC because the broad nature of the training set fits our purpose of designing a scoring system which is content agnostic and appropriate to users of any level of expertise. Questions and answers were generated by Amazon Mechanical Turk workers. Each source has an average of 11.15 questions and each question has four candidate answers. These candidate answers are labelled according to whether they are correct or incorrect, with 44.1% of answers being labelled as correct. Table 1 shows counts and wordcounts of sources, questions, and answers from the MultiRC dataset. During model development, we used a 70/15/15 train/validation/test split, splitting on sources so that no source is split between sets. This step is to protect against information leakage and ensure that the models will generalize outside of the training dataset.

Table 1: Descriptive Data on MultiRC Dataset

	n	Word Count Mean	Word Count SD
Sources	456	263.42	93.7
Questions	5,130	11.15	4.81
Answers	20,422	5.54	5.78

Although GPT performed well in generating short constructed response questions and answers, it is a proprietary tool owned by OpenAI, which raises concerns around privacy and interpretability [50] as well as cost if we were to use it to score constructed responses. To protect personally identifying information that may be inadvertently contained in a constructed response submitted by a user, we developed a scoring model that could be run locally instead of depending on GPT.

4.1.2. Model Selection

We used two methods for scoring the constructed responses. The first was a context-aware method that used the full context of the source, the question, and the answer to score the summary. The second was a context-independent method that involved distilling the relevant information in the source to a reference answer, then comparing the reference answer with the candidate answer provided by the student.

4.1.2.1. Context-aware models

For our first context-aware approach, we finetuned LLaMa-2-7b [40]. LlaMa, developed by Meta, has a similar architecture to GPT. Unlike GPT, however, information about LlaMa's training set and the weights of LlaMa itself were open-sourced by Meta. As a result, many innovative projects build off LlaMa. These include Stanford's Alpaca [46] and Vicuna [8]. We finetuned dolphin-llama2-7b using Low Rank Adaptation (LoRA), which is a parameter-efficient method of training large language models. Instead of updating and saving all model parameters, LoRA adapts the full parameter set to a lower rank matrix, thus allowing the training and fine-tuning of a large model with much lower compute requirements [21]. We finetuned lLaMa on

the training set to generate the token <<<TRUE>>> or <<<FALSE>>> based on the MultiRC sources, questions, and constructed responses using the following hyperparameters: per device train batch size = 4, gradient accumulation steps = 4, learning rate = 2e-4, max steps = 1,250, lora alpha = 16, lora dropout = 0.06, and r = 16.

For our second context-aware approach, we used Longformer [3], an encoder-only model. Encoder-only models such as Longformer utilize a special classification token. In the final layer of the encoder stack, the embedding associated with this token represents the entire document and can be used in downstream tasks, including binary classification. Longformer uses a sliding attention window to increase the max sequence length of the input while remaining computationally efficient [3]. This sliding attention window allowed us to include the MultiRC source and the question in the input without exceeding the maximum sequence length. We provided the MultiRC source, question, and answer to the Longformer model, separated by sep tokens (i.e. '</s>') similarly to in-context learning from previous studies [13, 35] to predict whether the MultiRC answer was correct. We finetuned Longformer for three epochs with a batch size of 8 and a learning rate of 3e-05 with accuracy as the reward function.

4.1.2.2. Context Independent Models

While the context-aware strategy has the advantage of using the entire context for prediction, it also requires a large amount of compute, translating into longer latency times. We thus tested more computationally efficient methods by using GPT-3.5 to distill the information from the MultiRC source and question into a reference correct answer, similar to a correct answer in an answer key. The reference correct answer is then compared to the MultiRC candidate answer using a much smaller encoder-only transformer and the classification token from the encoder model is used to predict whether the reference answer generated by GPT-3.5 is semantically similar to the MultiRC candidate answer. We tested two encoder-only reference models for this task – Masked and Permuted Language Modeling (MPNet) [43] and BiLingual Evaluation Understudy with Representations from Transformers (BLEURT) [39].

MPNet is an encoder-only model which unifies the masked language modeling training strategy employed by BERT with a permuted language modeling strategy employed by XLNet [51], leading to better performance in a range of downstream tasks [43]. The base model can then be finetuned on labelled data for specific tasks. As input to the model, we supplied the reference answers generated by GPT-3.5 from the MultiRC source and the candidate answers from the MultiRC dataset separated with a sep token (i.e. '</s>'). The labels used in finetuning were the 'is_correct' column of the MultiRC dataset. We trained mpnet-base for a binary classification task using accuracy as the reward function for 4 epochs with a batch size of 32 and a learning rate of 3e-05.

BLEURT [41] was designed as a deep-learning alternative to ROUGE [15] and BLEU [39] for predicting human judgements of the semantic similarity between two texts. BLEURT's training includes a pre-training step during which it is exposed to many synthetic pairs of candidate and reference texts to help it to better generalize across knowledge domains. We followed instructions from Huggingface (https://huggingface.co/Elron/bleurt-large-128) to further finetune the BLEURT-large model for classification with the MultiRC correct or incorrect answers as the labels. The input to the finetuning pipeline comprised the following structure: {"candidate": [student response], "reference": [gpt-generated response], "score": [score ∈ [0 … 1]]}. The score generated by BLEURT is continuous, ranging from 0 to 1 with 0 representing maximal dissimilarity and 1 representing maximal similarity. We finetuned the BLEURT model for 10 epochs at a batch size of 8 and a learning rate of 2e-5.

4.2. Results

Our context-aware models reported accuracies of .85 for the LLaMa2 model and .71 for the Longformer model. For our reference models, MPNET reported an accuracy of 0.81 and BLEURT reported an accuracy of 0.79. While the LLaMa2 model reported the highest accuracy, this came at the cost of much longer compute times. Although the context-independent models had slightly lower accuracy than the LLaMa2 model, they came with the benefit of lower compute requirements and faster processing. Computational efficiency is essential for automatic feedback systems because feedback latency is an important concern. Our presumption was that delaying feedback would lead to lower uptake of feedback and decrease motivation within iTELL [30]. As a result, we decided to use a reference model to provide feedback to users in iTELL about the accuracy of their constructed responses.

In a post-hoc study comparing the performance of MPNet and BLEURT when faced with adversarial attacks, we found that each model was weak against a different type of attack. For example, MPNet would sometimes score gibberish as correct, while BLEURT would inaccurately approve answers that were simply copied from the source. As a result, we decided to implement a consensus voting ensemble approach using both the MPNet and BLEURT models. For our consensus voting ensemble approach, if both models agree that the candidate response was incorrect, iTELL records a score of 0 and the student is told that their answer is incorrect and encouraged to revise their answer. If both models agree that the response is correct, iTELL records a score of 2 and the student is told that their answer is correct and they are encouraged to move to the next section of text. If the models disagree, iTELL records a score of 1 and the student is told that their answer is likely correct, and they are encouraged to revise their answer.

5. Efficacy Testing

After selecting the models for automatic constructed response question generation and constructed responses accuracy, we integrated these models into an iTELL deployment. Figure 1 shows the full pipeline for question generation and constructed response scoring. We used GPT-3.5 to generate questions and correct constructed responses for each language chunk. This happens only once during content preparation, with questions and constructed responses saved to a database. During use, the student submits their answer, and the answer is scored by two LLMs fine-tuned for the purpose. If the models agree that the candidate answer is correct, then the student receives a score of 2, if the models agree that the candidate answer is incorrect then the student receives a score of 0, while if the models disagree then the student receives a score of 1. We used the iTELL deployment to collect usage data from 98 participants to determine the efficacy of the automatic short constructed response scoring feature.

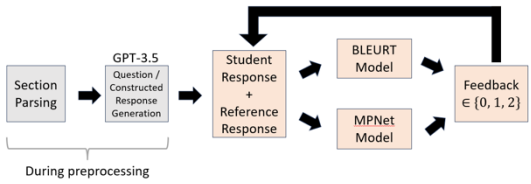

Figure 1: iTELL Question Generation and Constructed Response Scoring Pipeline

5.1. Method

5.1.1. Generating Intelligent Text

We adapted the first four chapters of the digital version of the *Think Python* [12] textbook for iTELL. As part of this process, each page of the text was divided into chunks. We used GPT-3.5 to generate questions and reference constructed responses for each chunk (excluding exercises, learning objectives, etc.). Questions and reference responses were checked by hand during the creation of the text. Table 2 shows the names of the chapters, the numbers of chunks in each chapter, and the mean and standard deviations of those chunk word counts.

Table 2: Descriptive Data on iTELL Intelligent Text

Page Name	Chunk (n)	Word Count mean	Word Count sd
Preface	3	36.67	50.58
The way of the program	7	267.14	165.9
Variables, expressions, etc.	10	223.1	137.8
Functions	12	214.75	93.7
Conditionals and recursion	12	203.33	59.1
Total	**44**	**209.73**	**117.3**

5.1.2. Integrating Constructed Responses

When a user begins reading a new page in iTELL, only the first chunk is visible while the rest of the page is blurred out. After reading the first chunk, the participants are prompted to click on a button to reveal the next chunk. Users continue in this manner until they complete the page. On one third of the chunks (determined at random), iTELL presents a constructed response question to the user. The user is required to attempt an answer before moving to the next chunk. The constructed responses are scored by iTELL's constructed response scoring model. If iTELL predicts that the user's candidate answer is incorrect, the text provides feedback and the user has the option to reveal the correct answer, skip the question, or revise and resubmit their answer. In addition to collecting data on success, we also collected data on user experience through thumbs up or thumbs down buttons in the constructed response scoring UI. Upon providing feedback, users were further invited (but not required) to select one or more tags to help explain their decision and write a short explanation. Figure 2 shows a screenshot of a question being presented within the iTELL intelligent text, with subsequent chunks blurred out until the user submits a response to the question.

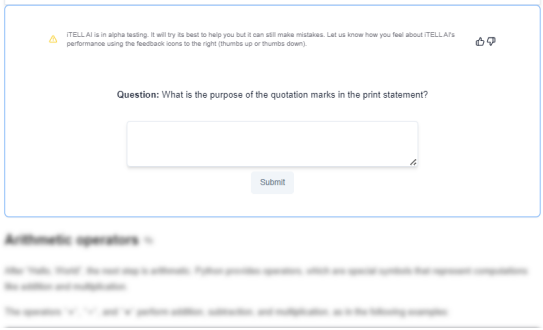

Figure 2: Presentation of Question in iTELL

5.1.3. Data Collection

After generating the iTELL deployment, we recruited participants for efficacy testing. The participants included 139 students in an introductory computer programming class that is part of an online computer science degree program. This study took place at the end of the semester so that participants were familiar with the content. Participants were offered extra credit for participating in the study. Of the students that participated, 98 indicated that they were over the age of 18 and consented to having their data used in this study.

Of these, 90 contributed further demographic information in this intake survey. All but one (98.9%) were between the ages of 18 and 24 years old and 84.4% (n = 76) were in the United States. All participants were proficient in English, with 87.8% (n = 79) being native or fully bilingual. The majority (80%, n = 72) had completed high school but not yet completed college. In terms of race and ethnicity, 42.2% (n =38) were Asian or Pacific Islander,

23.3% (n = 21) were White or Caucasian, while 11.1% (n = 10) were Black or African American and a further 11.1% (n = 10) were Hispanic or Latino.

After completing the iTELL intelligent text, participants were asked to complete an outtake survey to describe their experience of working with the text. In this survey, students were asked to rate how well the short constructed-response tool helped improve their learning, was easy to interact with, scored their answers accurately, and provided questions that were relevant to the subsection on a five-point Likert scale. Of the 98 participants who were over the age of 18 and who consented to allow their data to be used in the study, 82 provided responses on the outtake survey. A Cronbach's alpha conducted on the survey items in the outtake survey reported a score of $\alpha = 0.95$ (95% CI = [0.907, 0.955]) providing support for the reliability of the survey.

5.2. Results

The 98 participants produced 2,733 constructed responses. Figure 3 shows scores for each chunk in each page of the text. Scores of 2 (i.e., models agree that response is correct) were the most common, making up 64% (n = 1,749) of all scores. Scores of 0 (i.e. models agree that the response is incorrect) and 1 (i.e. models disagree) were rarer, making up 19.3% (n = 528) and 16.7% (n = 456) respectively.

To determine whether participants got better or worse at the constructed response task as they moved through the text, we performed a post-hoc linear regression on the proportion of each score {0, 1, 2} to the total number of responses for each subsection, regressing this proportion onto the chunk index. This analysis showed no significant effect on the proportions of 0s ($r = 0.23$, $p = 0.19$), 1s ($r = -0.17$, $p = 0.35$), or 2s ($r = -0.02$, $p = 0.92$) over time. This indicates that participants' performance on the constructed response task remained stable over time, with the notable exception of the first chunk, which had an exceptionally large number and proportion of 1s. This high proportion of 0s and 1s for the first question likely reflects students experimenting with the new question answering tool.

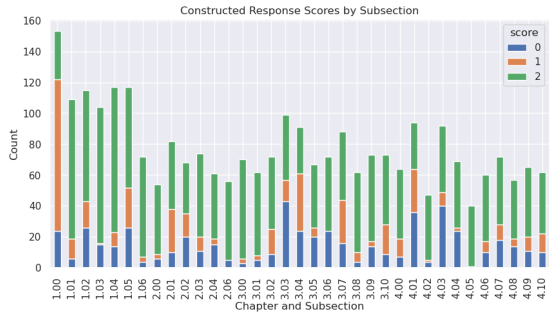

Figure 3: Results Across Pages and Chunks

Participants were invited to provide feedback on the constructed response task while they were responding to the questions by clicking on a thumbs up or thumbs down button and selecting up to three tags to explain why they provided that feedback. Thirty-eight students provided a total of 167 feedback responses on the constructed response items. Of these feedback responses, 75.4% (n = 126) were positive while 24.6% (n =41) were negative. Figure 4 shows counts of the tags provided by the participants. Of the 41 negative feedback items, inaccurate feedback was the main reason cited (n = 27). Only one response claimed that the question was harmful remarking that their answer was right, but the system marked it wrong, likely because of syntax. This participant left feedback that the system was "driving (them) crazy". This type of response was typical for the participants who reported negative experiences.

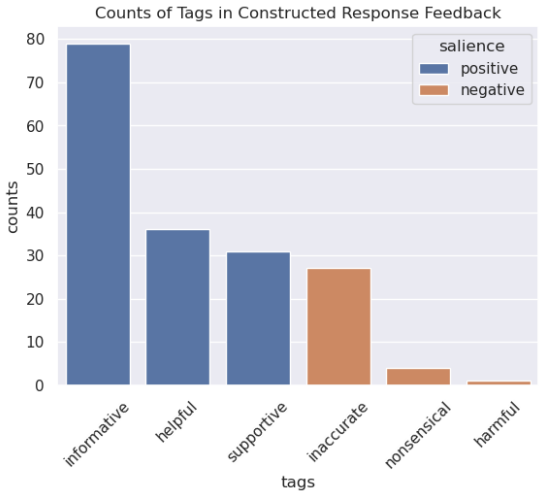

Figure 4: Counts of Feedback Tags

The outtake survey results generally included positive feedback for the constructed response tool. On the five-point Likert scale, 81.7% (n = 67) of respondents endorsed agreement or strong agreement to the statement that the short answer tasks helped them to improve their learning. Similarly, 84.1% (n = 69) of respondents agreed or strongly agreed that the constructed response task was easy to work with and 86.6% (n = 71) of respondents endorsed agreement or strong agreement to the statement that the questions were relevant to the subsection. A strong majority of participants (73.1%, n = 60) also agreed or strongly agreed that the model accurately scored their responses. Figure 5 shows participant responses for these questions.

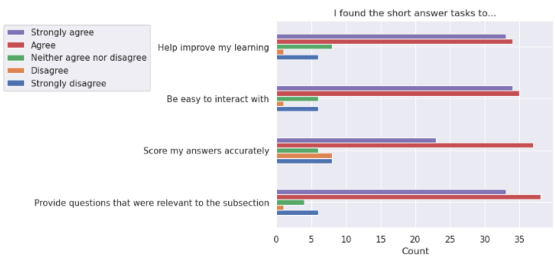

Figure 5: Feedback on Constructed Response Task

In addition to providing Likert-style responses, participants were also asked to provide written feedback on the outtake survey. Participants left a wide range of responses, including suggestions which will be used to further iterate the tool. These suggestions fell into five categories: comments on the accuracy of the feedback, suggestions for more complex questions to increase engagement, requests for greater frequency and variety of questions, requests for greater clarity and guidance in the feedback, and comments expressing satisfaction and effectiveness. Table 3 shows counts of each type of response, as well as a representative sample of each.

Table 3: Qualitative Written Feedback from Outtake Survey by Category

Category	Count	Representative Example
Accuracy	11	*sometimes the answers were marked as incorrect, even though they were correct*
Complexity	13	*I think they could be a little more engaging and not so simple*
Frequency	8	*I think there could be more of them within the textbook to help keep readers engaged.*
Feedback	14	*Give hints if someone gets it wrong*
Satisfaction	17	*These help me interact with the text and actually make me read the text.*
Other	18	*I have no opinion*
Total	**81**	

6. Discussion

This study investigated the process of designing a pipeline for automatic question generation and automatic constructed response scoring using LLMs in the context of iTELL, a framework for automatically converting static informative texts on any topic into a dynamic intelligent text. As part of iTELL, the LLM pipeline we designed had to be both automatic and domain agnostic. In this study we investigated whether GPT-3.5 could generate high-quality questions given a source context. Next, we experimented with several methods of scoring constructed responses from students, testing multiple finetuned LLMs. Finally, we tested our full question generation and constructed response scoring pipeline by inviting participants to use it as part of an iTELL intelligent text, collecting user feedback and suggestions.

In answer to the first research question, we showed that GPT-3.5 can generate high-quality questions and reference answers. In our trial of 30 sample chunks from the textbook on the Principles of Economics, GPT-3.5 generated sensible questions and correct answers that were relevant to the text. However, when prompted to generate questions of specific question types (i.e. recall, summary, inference), it performed no better than random chance at generating the appropriate question type. Instead, GPT-3.5 primarily generated recall questions. It is possible that specific question types could be elicited through different prompting strategies, or that new versions of GPT such as GPT-4 would perform better. This is a potential area for improvement in the system, a point which was echoed by the 13 students who suggested that the questions could be more complex and engaging.

For the second research question, regarding the best way to score the constructed responses, we tested four models using two scoring strategies. Both the lowest and the highest performance came from the context-aware strategy, which used information from the source, the question, and the candidate answer to determine whether the answer was correct or incorrect. Although LLaMa2 outperformed all other models, this high performance came at the cost of unacceptably increased feedback latency. The other context-aware model, Longformer, was much faster than LLaMa, but had the lowest performance of the models tested. The other two models used a reference strategy, comparing the candidate response to a pre-generated reference response which distills the context into a single phrase or sentence. These two models, MPNet and BLEURT, performed similarly. However, in post-hoc tests against adversarial attacks, they had different error profiles. As a result, and since inference with these models is fast, we decided to use both models in a consensus voting ensemble to determine whether the student answer is correct.

In answer to the third research question about user experience with the constructed response task, we found that user feedback was generally positive, with more than four fifths of users expressing agreement or strong agreement with the statement that the task helped them to improve their learning. In addition, 75.4% of the in-line feedback that we received as the students completed the task was positive. Of the negative feedback we received, the majority had to do with the accuracy of the feedback that the users received. In addition, we received many comments suggesting that we increase the frequency and complexity of the questions, as well as comments suggesting that we provide more clear formative feedback when users get an answer wrong.

7. Conclusion

This research indicates that the LLM pipeline for generating questions and scoring constructed responses to those questions is sufficient for integration into learning technologies like the iTELL framework. Answering automated questions through short constructed responses will contribute to the generation effect [4, 32, 34], enhancing interaction and learning.

Although the feedback we received from users was broadly positive, this study has several notable limitations. First, this study should be considered preliminary work that tested the potential for LLMs to generate and evaluate constructed response items in a digital text. Therefore, we

need to further evaluate the impact of this system on learning outcomes. In future work, we plan to address the impact of constructed response tasks through randomized controlled trials which will compare learning in versions of iTELL with and without constructed response items (A/B testing). This research will help explain the impact of the constructed response activity on learning and reading comprehension.

Second, users expressed twin concerns about the feedback and accuracy of the model. These concerns are related because the model is not capable of clearly explaining why a given constructed response failed. Future research may employ generative LLMs to help users move step by step from an incorrect answer to a correct answer, perhaps by giving hints or suggestions using chain-of-thought prompting [49]. Providing an explanation when the model judges an answer to be wrong may help improve the user experience of model accuracy in addition to providing clearer feedback.

Third, iTELL currently deploys the questions randomly for one third of the chunks. While this strategy appears to work well, it is unlikely to be the optimal strategy. Other ways to deploy questions include analyzing reading behavior, for example, requiring users to answer questions about subsections that they skip or read through quickly. Alternatively, we could assess the reading difficulty of the chunks themselves by comparing their similarity with the summaries that students write at the end of each chapter, assigning readers to answer questions about chunks that are more often ignored. Finding an optimal question deployment strategy is a subject for future research.

Finally, the current scoring models simply compare the candidate answer from the student to a reference answer generated by a model, similarly to how a teacher might use an answer key to score responses. This may work for lower order thinking skills such as recall and summarization, but answers to questions that require higher order thinking skills such as inference, bridging, or logic may come from outside the text [33]. It is likely that a different scoring approach would be needed for higher order question types.

Overall, our research on question generation and constructed response scoring indicates the potential to develop automatic NLP pipelines to develop text-based questions and score resulting constructed responses. Such approaches should allow for enhanced assessment of reading comprehension, reading development, and knowledge acquisition. As LLMs become more prevalent, more powerful, and more practical, we can expect that personalized learning tools such as iTELL will become more and more ubiquitous. We hope that these developments will lead to improved outcomes for students and adult learners within advanced learning platforms.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant 2112532 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

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

[2] Begg, I., Snider, A., Foley, F. and Goddard, R. 1989. The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15, 5 (Sep. 1989), 977–989. DOI:https://doi.org/10.1037/0278-7393.15.5.977.

[3] Beltagy, I., Peters, M.E. and Cohan, A. 2020. Longformer: the long-document transformer. (2020). DOI:https://doi.org/10.48550/ARXIV.2004.05150.

[4] Bertsch, S., Pesta, B.J., Wiscott, R. and McDaniel, M.A. 2007. The generation effect: A meta-analytic review. *Memory & Cognition*. 35, 2 (Mar. 2007), 201–210. DOI:https://doi.org/10.3758/BF03193441.

[5] Blything, L.P., Hardie, A. and Cain, K. 2020. Question Asking During Reading Comprehension Instruction: A Corpus Study of How Question Type Influences the Linguistic Complexity of Primary School Students' Responses. *Reading Research Quarterly*. 55, 3 (Jul. 2020), 443–472. DOI:https://doi.org/10.1002/rrq.279.

[6] Brusilovsky, P., Sosnovsky, S. and Thaker, K. 2022. The return of intelligent textbooks. *AI Magazine*. 43, 3 (Sep. 2022), 337–340. DOI:https://doi.org/10.1002/aaai.12061.

[7] Chen, O., Kalyuga, S. and Sweller, J. 2015. The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology*. 107, 3 (Aug. 2015), 689–704. DOI:https://doi.org/10.1037/edu0000018.

[8] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y. and Gonzalez, J.E. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90\%* chatgpt quality.

[9] Crossley, S., Kyle, K., Davenport, J. and Danielle S., M. 2016. Automatic Assessment of Constructed Response Data in a Chemistry Tutor. *International Educational Data Mining Society* (Raleigh, NC., 2016).

[10] Crutcher, R.J. and Healy, A.F. 1989. Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15, 4 (Jul. 1989), 669–675. DOI:https://doi.org/10.1037/0278-7393.15.4.669.

[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.

[12] Downey, A.B. 2015. *Think Python*. O'Reilly.

[13] Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R. and Lan, A. 2022. Automated Scoring for Reading Comprehension via In-context BERT Tuning. *Artificial Intelligence in Education*. M.M. Rodrigo,

N. Matsuda, A.I. Cristea, and V. Dimitrova, eds. Springer International Publishing. 691–697.

[14] Gamage, D., Staubitz, T. and Whiting, M. 2021. Peer assessment in MOOCs: Systematic literature review. *Distance Education*. 42, 2 (Apr. 2021), 268–289. DOI:https://doi.org/10.1080/01587919.2021.1911626.

[15] Ganesan, K. 2018. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. (2018). DOI:https://doi.org/10.48550/ARXIV.1803.01937.

[16] Giannakopoulos, K.L., McCurdy, M.P., Sklenar, A.M., Frankenstein, A.N., Levy, P.U. and Leshikar, E.D. 2021. Less Constrained Practice Tests Enhance the Testing Effect for Item Memory but Not Context Memory. *The American Journal of Psychology*. 134, 3 (Oct. 2021), 321–332. DOI:https://doi.org/10.5406/amerjpsyc.134.3.0321.

[17] Gillioz, A., Casas, J., Mugellini, E. and Khaled, O.A. 2020. Overview of the Transformer-based Models for NLP Tasks. (Sep. 2020), 179–183.

[18] Greenlaw, S. and Shapiro, D. 2017. *Principles of Macroeconomics*. OpenStax.

[19] He, P., Liu, X., Gao, J. and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv.

[20] Hift, R.J. 2014. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education*. 14, 1 (Dec. 2014), 249. DOI:https://doi.org/10.1186/s12909-014-0249-2.

[21] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. (2021). DOI:https://doi.org/10.48550/ARXIV.2106.09685.

[22] Jacoby, L.L. 1978. On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*. 17, 6 (Dec. 1978), 649–667. DOI:https://doi.org/10.1016/S0022-5371(78)90393-6.

[23] Jay Robert Campbell 1999. Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. *ProQuest Dissertations & Theses Global*. (1999).

[24] Jung, Y.J., Crossley, S. and McNamara, D. 2019. Predicting Second Language Writing Proficiency in Learner Texts Using Computational Tools. *The Journal of AsiaTEFL*. 16, 1 (Mar. 2019), 37–52. DOI:https://doi.org/10.18823/asiatefl.2019.16.1.3.37.

[25] Kalyan, K.S. 2024. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*. 6, (Mar. 2024), 100048. DOI:https://doi.org/10.1016/j.nlp.2023.100048.

[26] Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S. and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, 2018), 252–262.

[27] Koubaa, A. 2023. *GPT-4 vs. GPT-3.5: A Concise Showdown*. ENGINEERING.

[28] Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*. 30, 1 (Mar. 2020), 121–204. DOI:https://doi.org/10.1007/s40593-019-00186-y.

[29] Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J. and Kim, H. 2023. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*. (Oct. 2023). DOI:https://doi.org/10.1007/s10639-023-12249-8.

[30] Maier, U., Wolf, N. and Randler, C. 2016. Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*. 95, (Apr. 2016), 85–98. DOI:https://doi.org/10.1016/j.compedu.2015.12.002.

[31] McCarthy, K.S., Allen, L.K. and Hinze, S.R. 2020. Predicting Reading Comprehension from Constructed Responses: Explanatory Retrievals as Stealth Assessment. *Artificial Intelligence in Education*. I.I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, eds. Springer International Publishing. 197–202.

[32] McCurdy, M.P., Viechtbauer, W., Sklenar, A.M., Frankenstein, A.N. and Leshikar, E.D. 2020. Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*. 27, 6 (Dec. 2020), 1139–1165. DOI:https://doi.org/10.3758/s13423-020-01762-3.

[33] McNamara, D.S. 2017. Self-Explanation and Reading Strategy Training (SERT) Improves Low-Knowledge Students' Science Course Performance. *Discourse Processes*. 54, 7 (Oct. 2017), 479–492. DOI:https://doi.org/10.1080/0163853X.2015.1101328.

[34] McNamara, D.S. and Healy, A.F. 2000. A Procedural Explanation of the Generation Effect for Simple and Difficult Multiplication Problems and Answers. *Journal of Memory and Language*. 43, 4 (Nov. 2000), 652–679. DOI:https://doi.org/10.1006/jmla.2000.2720.

[35] Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M. and McNamara, D. 2024. Formative Feedback on Student-Authored Summaries in Intelligent Textbooks Using Large Language Models. *International Journal of Artificial Intelligence in Education*. (Mar. 2024). DOI:https://doi.org/10.1007/s40593-024-00395-0.

[36] Morris, W., Crossley, S., Holmes, L., Ou, C., McNamara, D. and Dascalu, M. 2023. Using Large Language Models to Provide Formative Feedback in Intelligent Textbooks. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O.C. Santos, eds. Springer Nature Switzerland. 484–489.

[37] Morris, W., Holmes, L., Chui, J.S. and Crossley, S. Under review. Automated Scoring of Constructed

Response Items in Math Assessment Using Large Language Models.

[38] Ouyang, L. et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* (2022).

[39] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. 2001. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (Philadelphia, Pennsylvania, 2001), 311.

[40] Roumeliotis, K.I., Tselikas, N.D. and Nasiopoulos, D.K. 2023. *Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model*. Computer Science and Mathematics.

[41] Sellam, T., Das, D. and Parikh, A.P. 2020. BLEURT: Learning Robust Metrics for Text Generation. (2020). DOI:https://doi.org/10.48550/ARXIV.2004.04696.

[42] Slamecka, N.J. and Graf, P. 1978. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*. 4, 6 (Nov. 1978), 592–604. DOI:https://doi.org/10.1037/0278-7393.4.6.592.

[43] Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *Advances in neural information processing systems*. 33, (2020).

[44] Sukkarieh, J., Pulman, S. and Raikes, N. 2003. *Automarking: using computational linguistics to score short free-text responses*.

[45] Sukkarieh, J.Z. and Blackmore, J. 2009. c-rater:Automatic Content Scoring for Short Constructed Responses. *Flairs Conference*. (2009).

[46] Taori, R., Gulrajani, I., Yang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T. 2023. Stanford Alpaca: An Instruction-following LLaMa model.

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. Attention Is All You Need. (2017). DOI:https://doi.org/10.48550/ARXIV.1706.03762.

[48] Wang, Z., Valdez, J., Basu Mallick, D. and Baraniuk, R.G. 2022. Towards Human-Like Educational Question Generation with Large Language Models. *Artificial Intelligence in Education*. M.M. Rodrigo, N. Matsuda, A.I. Cristea, and V. Dimitrova, eds. Springer International Publishing. 153–166.

[49] Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, brian, Xia, F., Chi, E., Le, Q.V. and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* (2022), 24824–24837.

[50] Wu, X., Duan, R. and Ni, J. 2023. Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence*. (Oct. 2023), S2949715923000707.
DOI:https://doi.org/10.1016/j.jiixd.2023.10.007.

[51] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems* (2019).