Generating Reading Assessment Passages Using a Large Language Model

Ummugul Bezirhan
International Study Center
Boston College
bezirhan@bc.edu

Matthias von Davier
Boston College
vondavim@bc.edu

Abstract

The growing demand for high-quality items in computer-based assessments has made the item creation process costly and labor-intensive, relying heavily on human expertise. While automated item generation has been around, large language models can enhance efficiency and quality. In this study, we explored the use of GPT family models to generate reading passages for the Progress in International Reading Literacy Study (PIRLS). Creating passages for 4th graders requires careful attention to complexity, engagement, and relevance. By using well-designed prompts, we generated multiple passages and selected those closely matching original texts based on Lexile scores. All AI-generated passages along with original passages are evaluated by human judges according to their coherence, appropriateness to 4th graders, and readability.

1 Introduction

The technological innovations in all aspects of test development facilitate efficient test practices and a more well-rounded information retrieval from the data compared to traditional paper-pencil assessments. Consequentially, the integration of technology in computer-based assessments (CBA) increases the demand for more frequent administration and rapid and efficient production of high-quality content-specific innovative items. The greater selection of item types presented to an examinee in a continual fashion necessitates a more streamlined item development process. Conventional item development is one of the most expensive, time-consuming, and labor-intensive parts of assessment development because the process heavily depends on human content specialists. Human subject experts write each item individually, then each item is reviewed, edited, and revised by a group of experts until it meets predefined quality control standards (Haladyna, 2013). Therefore, the subjectivity of traditional item writing is often compromised by the subject experts' qualifications and understanding of the specific content area. The other issues with the traditional item development process are its lack of efficiency and scalability. Automated Item Generation (AIG) is proposed to address the limitations associated with conventional item development by utilizing cognitive and psychometric theories with the help of computer technology to generate items [Gierl and Haladyna, 2013].

A considerable amount of literature has already been published on AIG in educational measurement. These studies typically utilized the template-based approach [Gierl and Lai, 2013] for generating assessment items. This approach usually uses a three-stage procedure to automatically capture necessary information and features to produce multiple-choice items (MCQs). First, content experts build a cognitive model that defines the knowledge, skills, and content that are needed to process and solve given questions. In the second stage, content experts create an item model to highlight the parts and content of the assessment task that can be manipulated to create new items. The item template is a prototypical representation of a test item that informs the automated item generation process. In the

NeurIPS 2024 Workshop on Large Foundation Models for Educational Assessment (FM-Assess)

last step, a computer algorithm utilizes information from both cognitive - and item-model to generate new items.

While this approach reduced the cost and time associated with traditional item writing, it still suffers the reliance on generating clones of narrowly defined item types by only manipulating limited task components of certain items to derive item templates [von Davier, 2018]. Another important limitation of templated-based approaches is the human expert associated cost. The automation process does not start until after considerable groundwork takes place by content experts. Kosh et al. [2019] argued that the cost-effectiveness of template-based AIG depends on most of the items being generated to belong in the same content area and tests with a limited number of skills that can be modeled with a single cognitive model.

Another approach that has been explored in the AIG framework focused on generating items without any prespecified templates and human intervention by mostly utilizing Natural Language Processing (NLP) techniques [Shin, 2021]. Specifically, part of speech tagging, topic modeling, and noun phrase extraction have been explored in question generation [Azevedo et al., 2020, Flor and Riordan, 2018, Mazidi, 2018] In his inaugural paper, von Davier [2018] demonstrated the use of a novel neural network approach, long short-term memory-based recurrent neural networks (LSTM-RNN) to generate personality items. With these methods, the textual information is extracted and modeled from a collection of documents which replaced the manual construction of cognitive and content models to generate items.

With the advances in the field of NLP, large transformer-based language models in other words self-attention a such as Bidirectional Encoder Representations from Transformers (BERT; Devlin [2018]) and the Generative Pretrained Transformer (GPT; Radford et al. [2019]) often approach human-level performance in diverse language tasks. The earlier version of the GPT model, GPT-2, was released by OpenAI in 2019 and it is subsequently utilized for medical education [von Davier, 2019] and personality question generation [Hommel et al., 2022]. Despite the groundbreaking performance of GPT-2 compared to other language models, it was not equipped to handle more specialized tasks such as storytelling and constructing complex language formations. For a specialized task, GPT-2 required sufficient pre-training to be able to generate appropriate responses.

OpenAI released Generative Pre-trained Transformer 3 (GPT-3) which excels at few-shot learning, which means it can be given a small number of representative samples of text to complete a task without any pre-training, such as text generation [Brown et al., 2020]. The goal of this research is to utilize GPT family models to generate reading comprehension items in the context of an international large-scale assessment.

2 Method

Progress in International Reading Literacy Study (PIRLS) is an international study of primary school students' reading skills and is administered every five years. The assessment consists of a battery of tasks, including literary and informational passages with accompanying multiple-choice and open-ended questions. Students' reading literacy is assessed with the passages that are drawn from a wide range of genres either in fiction or non-fiction. We utilized GPT models to generate passages in similar genres as those in the PIRLS assessment, and human judges examined the passages according to their coherence and appropriateness.

In this study, we utilized OpenAI's text-davinci-002 model, also referred to as InstructGPT, released in January 2022 as an updated and fine-tuned version of GPT-3 to generate reading passages for PIRLS, which assesses reading literacy in 4th graders across two text types: literary and informational. Following the PIRLS framework, we constructed prompts for both types using a pool of 24 released passages from PIRLS cycles (2001–2016). Three restricted-use PIRLS passages—two informational and one literary—served as a reference for generating new passages: Antarctica: Land of Ice, Ants, and Brave Charlotte.

We used Python to interact with GPT-3's text completion API, generating passages through well designed prompts. Prompts were tested in both zero-shot and one-shot learning settings. In one-shot learning, we included a demonstration passage alongside the prompt to guide the model, while in zero-shot learning, only an instruction was provided. For both methods, we included details about the target audience's age to ensure the passages were appropriate for 4th graders. This additional information is provided to the model to ensure that the generated passages are age-appropriate, as they are intended for use in a fourth grade assessment. An example prompt for each scenario is given in Appendix. We first generated outputs with one-shot learning (Figure A.1), after determining the topic of the story from the first set of outputs, we utilized that in the prompt of both one-shot (Figure A.2) and zero-shot (Figure A.3) learning for more detailed prompts.

We varied the model's temperature setting (t = 0.5, 0.7, 0.9) to control the creativity and diversity of the outputs. Lower temperatures generated more predictable and consistent text, while higher temperatures produced more creative and diverse outputs. Ten replications per temperature setting were generated for each task. After generating the passages, the selection mechanism involved first calculating the text difficulty score using an online text difficulty analyzer [Cathoven, 2023] associated with each generated story. This score is similar to a Lexile score [Stenner, 2022] that indicates the level of reading difficulty by combining measures of semantics and syntax that are represented by word frequency counts and sentence length, respectively. After calculating the text difficulty scores for both the original PIRLS passages and the generated passages, only the generated passages that fell within one standard deviation of the original passages' text difficulty scores were selected for further evaluation.

Three human editors reviewed the selected passages for grammar, coherence, and factual accuracy, especially in informational texts. To assess the appropriateness of these passages, we conducted an online survey with 150 participants who work in the education sector. Participants used a 4-point Likert scale to rate the passages. After giving each passage following questions were asked: "The story is written at an adequate reading level for a fourth grader," "The story is written in a coherent manner," "Children will be able to identify the main topic of the story," "There are confusing or distracting elements in the story," and "This story can engage children to answer questions." We recruited 150 participants (50 for each survey) through Amazon Mechanical Turk. We also set up additional qualifications for the participant selection. To be a part of the study, participants must have a Bachelor's degree and be working in the education industry. Additionally, we used different measures to identify and exclude potential inattentive users and non-respondents (bots) in Amazon MTurk. To identify bots, we increased the time between Human intelligence task (HIT) completion and auto-approval to examine the data before approving or rejecting the HITs. Moreover, we rejected HITs with unreasonable response times, this was determined using median absolute deviation statistic on the overall completion time. The last measure was to incorporate an attention question in each survey to sort out careless respondents. This resulted in a final sample of 50 respondents for each survey.

3 Results

For the informational PIRLS passages, Antarctica: Land of Ice and Ants, we used both one-shot and zero-shot learning, incorporating an additional age/grade indicator. This process generated passages such as The Amazon: Green Lungs of the Planet and Bees. A total of 160 passages were generated, with 40 created for each condition. Text difficulty scores were calculated for all outputs. Initially, we applied one-shot prompt allowing GPT-3 to generate passages similar to the example provided. The first batch of 10 replications was used to determine the topic and subsequently discarded. GPT-3 consistently generated passages on similar topics such as for Ants, most passages centered around Bees.

Figure 1 shows the distribution of text difficulty scores for the Bees passage, generated using both one-shot and zero-shot learning, with and without grade information. The Bees passage is based on the PIRLS Ants passage, which had a text difficulty score of 560, represented by the horizontal line in the figure. Among all prompt types, one-shot learning with grade information produced passages with lower text difficulty scores compared to the others. While the inclusion of grade information tended to reduce the difficulty, no clear pattern was observed regarding the variance across prompt types.

The other informational passage "The Amazon: Green Lungs of the Planet" was prompted using the PIRLS "Antarctica: Land of Ice" passage as an example. It has to be pointed out that these generated passages are not simple copies where one word is replaced by another word to vary content. Unlike traditional AIG approaches [Gierl et al., 2020], the GPT-generated passages are based on priming a large language model with a context, and then have the model generate an independent

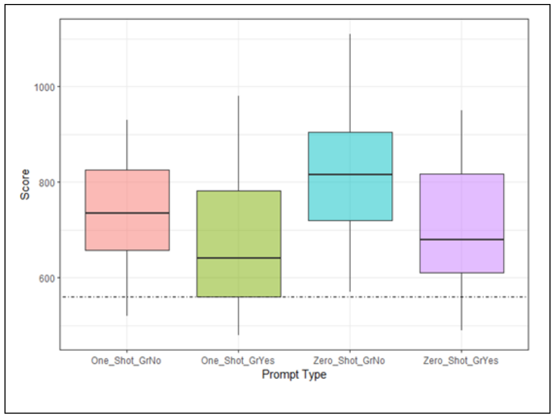

Figure 1: Text difficulty scores for generated "Bees" passage under different prompting conditions.

text, inspired by requesting a type of text, a topic, or by providing an example. The calculated text difficulty scores are presented in Figure 2, the horizontal line shows the score for the "Antarctica: Land of Ice" passage. For both one-shot and zero-shot prompting conditions, inclusion of grade/age information reduced the text difficulty score for the generated passages.

For the literary passage generation, the prompting approach differed slightly from the informational one. As with the informational passages, the first step was to generate story ideas based on an existing PIRLS passage. We used Brave Charlotte, a story about a brave sheep helping a shepherd in a tough situation, as the initial prompt. GPT-3 produced various storylines around themes such as friendship, kindness, and community, which aligned with the general tone of the original story. From these ideas, we developed Coco the Rabbit. However, generating longer literary passages posed a challenge due to GPT-3's tendency to produce premature stopping points. To address this, we adopted a stepwise approach, using previous outputs as prompts along with additional instructions to successfully generate full-length stories.

The finalized stories were included in the data collection alongside the PIRLS passages. To prevent participants from recognizing similarities between prompted and generated texts, we paired the Bees passage with Antarctica: Land of Ice and The Amazon: Green Lungs of the Planet with Ants for informational texts. For literary passages, we paired the generated and original stories together since they had different characters and storylines.

We used a four-category Likert scale to assess participant attitudes toward each passage, gathering responses from 50 participants per passage. Figure 3 shows the distribution of responses for the

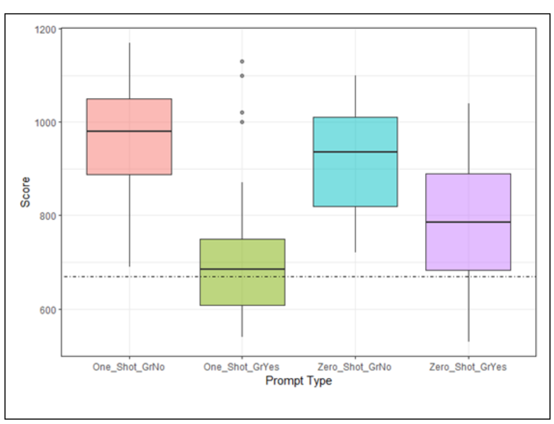

Figure 2: Text difficulty scores for generated "The Amazon: Green Lungs of the Planet" passage under different prompting conditions.

GPT-3-generated Bees passage and the original PIRLS Ants passage. Overall, 92% of participants agreed or strongly agreed that the GPT-3 passage was appropriate, while 96% felt the same about the original. Responses regarding engagement were similarly positive, with the original passage scoring 86% and the AI-generated passage 84%. The largest gap was in coherence, where 94% of participants found the original passage coherent, compared to 84% for the AI-generated passage.

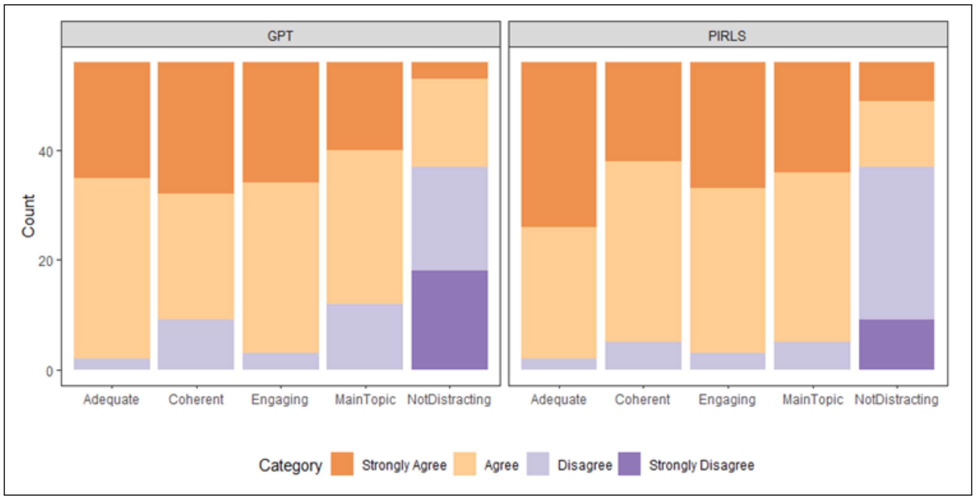

Figure 3: GPT generated “Bees” and PIRLS original “Ants” passage survey results

Comparable results were observed in the comparison between Antarctica: Land of Ice and The Amazon: Green Lungs of the Planet. While the overall levels of agreement were consistent between the AI-generated and original passages, individual response categories showed some variations for certain questions. The distribution of responses and level of agreement and disagreement among participants for these passages are given in Figure 4. Overall, the agreement and disagreement levels aligned well between the AI-generated and the original human-generated PIRLS passages across all questions. However, we observed a larger difference between the individual response categories for certain questions.

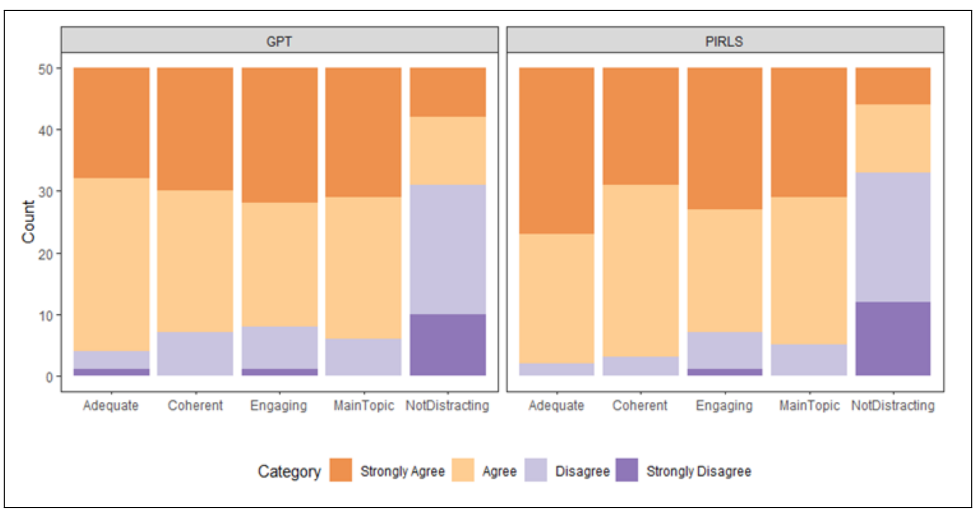

Figure 4: GPT generated “Amazons” and PIRLS original “Antarctica” passage survey results

For both “Amazons” and “Antarctica” passages, 97% of participants agreed that the passages were adequate for fourth grade readers. However, the strongly agree category was higher for the original PIRLS passage (54%) compared to the AI-generated one (38%). For coherence, the positive attitude towards the original PIRLS passage (91%) was higher compared to GPT generated passage (84%), this suggests that human-generated passages may have been more logically structured and easier to follow than the GPT generated one. However, more judges strongly agreed (43%) with the statement for the GPT generated passage compared to the original PIRLS passage (32%), when we look at the individual categories for coherence. Moreover, identifying the passage's main topic was deemed harder for GPT generated passage, with 79% of the participants agreeing with the statement, compared to the agreement with the human-written passage (91%). Finally, the overall agreement on the passage not being distracting was similar for both passages, but more judges had a stronger evaluation towards GPT generated passage (32%) being more distracting than the original PIRLS passage (16%). However, when taking strongly agree and agree categories together, there was no difference observed between the general level of agreement for AI generated (66%) vs. human generated passage (66%).

Lastly, Figure 5 displays the level of agreement and disagreement between "Coco the Rabbit" and "Brave Charlotte" stories. A somewhat similar pattern in the agreement was also observed with these literary passages, with GPT generated passage being slightly more adequate and engaging and less distracting compared to the original PIRLS passage. For the passage being adequate for the fourth graders, judges agreement was higher for AI-generated passage (96%) compared to the original PIRLS passage (80%). A similar pattern was observed for the engagement, 94% of the participants agreed with the AI-generated passage to be engaging, whereas only 74% of the participants agreed about the same statement for the original passage. Lastly, about 16% more people found the human written passage more distracting compared to the GPT-generated passage.

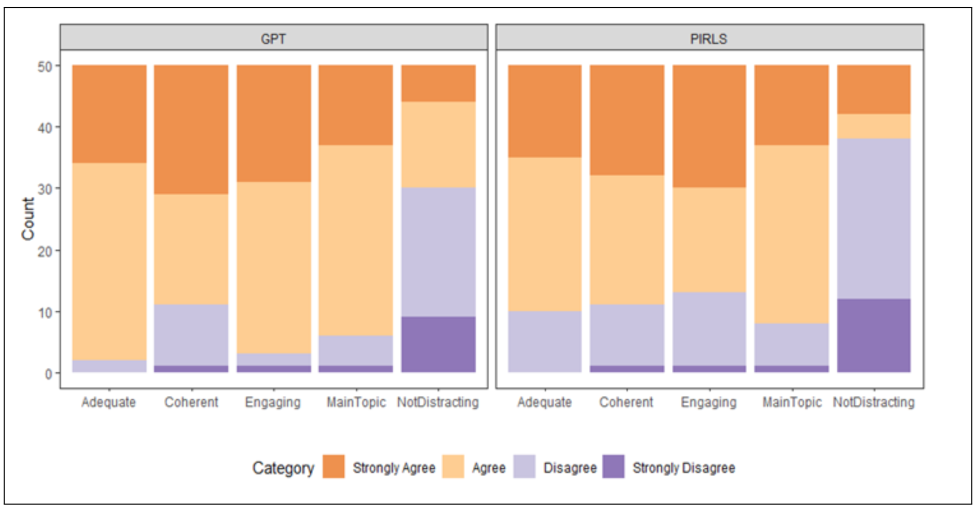

Figure 5: GPT generated "Coco the Rabbit" and PIRLS original "Brave Charlotte" passage survey results

4 Conclusion

This study explored the use of GPT-3 in generating literary and informational passages for an international large-scale reading assessment, specifically PIRLS. We experimented with different prompt designs, incorporating audience age information, to generate passages that match the structure and difficulty of original PIRLS texts.

Our findings indicate that one-shot learning with detailed prompts produced the best results in terms of text difficulty alignment with original passages. This supports prior research showing that GPT-3 performs better with more examples and clear instructions. The results also underscore the importance of direct task specification in guiding GPT to produce relevant and accurate content. In particular, GPT was able to generate passages that closely mirrored the original PIRLS texts in terms of length, vocabulary, and difficulty.

However, the analysis revealed that GPT-3-generated informational passages were sometimes more distracting and less coherent than the human-authored PIRLS passages, likely due to the absence of intentional organization. In contrast, the literary passages generated through iterative prompting were found to be less distracting and more engaging than their PIRLS counterparts, suggesting that GPT-3's storytelling capabilities can benefit from more dynamic prompt engineering.

Despite promising results, there are limitations that future research should address. While this study examined some prompt design strategies, further exploration of both manual and automated prompt engineering techniques could optimize the generation of age-appropriate reading passages. Finally, the small sample size used in the empirical analysis limits the generalizability of the findings.

It is also important to address the veracity and fact validation practices while utilizing large language models to create contextual texts. GPT family models have the ability to convince users and induce trust in the output of these models due to the coherency, naturalness and human-like quality of its responses despite potential inaccuracies [Sison et al., 2024]. Thus, it is imperative to incorporate fact checking mechanisms within models and systems utilizing this technology. One potential way to achieve this is to develop human-in-the-loop mechanisms, similar to the approach employed in this paper, where human agents participate in verifying the reliability and accuracy of the generated text. Additionally, exploring novel techniques such as integrating real time fact checking databases into the models and developing systems that align with the principles of a Human-Centered AI framework [Shneiderman, 2020] can contribute to improving the reliability and accuracy of AI-generated content.

Overall, this research demonstrates that GPT family models could be effectively utilized for automated passage generation in the context of a large-scale reading assessment. Considering the high costs and significant time investment associated with human-authored assessment development and the copyright concerns that often arise, large language models present a promising opportunity to streamline and enhance current practices in assessment development.

References

Pedro Azevedo, Bernardo Leite, Henrique Lopes Cardoso, Daniel Castro Silva, and Luís Paulo Reis. Exploring nlp and information extraction to jointly address question generation and answering. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 396–407. Springer, 2020.

T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners advances in neural information processing systems 33. 2020.

Cathoven. Text difficulty analyzer, 2023. URL `https://www.cathoven.com/en/freetext-difficulty-analyzer/`.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Michael Flor and Brian Riordan. A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 254–263, 2018.

Mark J Gierl and Thomas M Haladyna. *Automatic item generation: Theory and practice*. Routledge, 2013.

Mark J Gierl and Hollis Lai. Evaluating the quality of medical multiple-choice items created with automated processes. *Medical education*, 47(7):726–733, 2013.

Mark J Gierl, Hollis Lai, and Donna Matovinovic. Augmented intelligence and the future of item development. *Application of artificial intelligence to assessment*, pages 1–25, 2020.

Björn E Hommel, Franz-Josef M Wollang, Veronika Kotova, Hannes Zacher, and Stefan C Schmukle. Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2):749–772, 2022.

Audra E Kosh, Mary Ann Simpson, Lisa Bickel, Mark Kellogg, and Ellie Sanford-Moore. A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, 38 (1):48–53, 2019.

Karen Mazidi. Automatic question generation from passages. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*, pages 655–665. Springer, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Eunjin Shin. Automated item generation by combining the non-template and template-based approaches to generate reading inference test items. 2021.

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

Alejo Jose G Sison, Marco Tulio Daza, Roberto Gozalo-Brizuela, and Eduardo C Garrido-Merchán. Chatgpt: More than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (hcai) perspective. *International Journal of Human–Computer Interaction*, 40(17):4853–4872, 2024.

A Jackson Stenner. Measuring reading comprehension with the lexile framework. In *Explanatory models, unit standards, and personalized learning in educational measurement: Selected papers by A. Jackson Stenner*, pages 63–88. Springer, 2022.

Matthias von Davier. Automated item generation with recurrent neural networks. *psychometrika*, 83 (4):847–857, 2018.

Matthias von Davier. Training optimus prime, md: Generating medical certification items by fine-tuning openai's gpt2 transformer model. *arXiv preprint arXiv:1908.08594*, 2019.

A Appendix

Example Prompts are given below.

```
This is an informative story generator.
The story should have multiple parts and the sections should be informative
and engaging [for a 10-year-old]. An example story is given below.

Story: Ants
Small and Strong
Lift up a rock, and a family of ants might be crawling there.
Ants are small insects, but they are very strong. Ants have six strong legs
that help them carry big loads such as sticks and other insects. They can
lift 20 times their own body weight.

Building a Home
Most ants live in nests in the ground. Each nest is like an underground city.
It has rooms, called chambers, where the ants live and work.
The chambers are connected by tunnels.
..........
```

Figure A.1: Initial Prompt for Bees Passage

```
This is an informative story generator.
Generate an informative story about Bees [for a 10-year-old]. It includes sections about
bees' body, their honey production, social life and importance to ecosystem.
The sections should be informative and engaging [for a 10-year-old].

Story: Ants
Small and Strong
Lift up a rock, and a family of ants might be crawling there.
Ants are small insects, but they are very strong. Ants have six strong legs
that help them carry big loads such as sticks and other insects. They can
lift 20 times their own body weight.

Building a Home
Most ants live in nests in the ground. Each nest is like an underground city.
It has rooms, called chambers, where the ants live and work.
The chambers are connected by tunnels.
..........
```

Figure A.2: One-shot prompt for Bees Passage

```
This is an informative story generator.
Generate an informative story about Bees [for a 10-year-old]. It includes sections about
bees' body, their honey production, social life and importance to ecosystem.
The sections should be informative and engaging [for a 10-year-old].
```

Figure A.3: Zero prompt for Bees Passage

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The data, original stories and the code could be obtained upon request.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data, original stories and the code could be obtained upon request.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The full paper specifies all settings, prompts and parameters that were used in the model usage.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper clearly states the origins of the original stories (PIRLS) used to generate passages.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Justification: The full paper includes the newly generated stories, if those are considered assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes information on human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification:Justification: Study do not pose any risk to participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.